Data Visualization of Arsenal F.C. 2021-2022 Season

Will Rauen

2024-02-02

Setup

We will be using the tidyverse and ggplot2 packages in R for this assignments data analysis and visualization.

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.4     v purrr   1.0.2
## v tibble  3.2.1     v dplyr   1.1.4
## v tidyr   1.3.1     v stringr 1.5.1
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
require(ggplot2)
require(gridExtra)
```

```
## Loading required package: gridExtra
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 90), tidy = TRUE)
```

Data

```
df <- read.csv("/Users/williamrauen/Desktop/DS 3100/Assignments/Data/2021-2022.csv")

dim(df)  #looks at the dimension of the dataframe
```

```
## [1] 380 106
```

The data used in this dataset comes from the 2021-2022 season of England's Premier League and comes from kaggle: https://www.kaggle.com/datasets/saife245/english-premier-league/data.

The Premier League is the top division of English football in which there are 20 teams who play each other twice a season where each team plays once in their home stadium. Each row of the data represents one game of the premier league totaling to a total of 380 games. Of the variables included in the data set, there are

stats pertaining to number of goals scored by each team, who was the match referee, scores at half-time and full-time, number of shots, number of fouls, and a few other stats. Something of note in this dataset is that it comes with no missing variables, so data cleaning related to absent values will not be necessary.

Research Question

In the 2021-2022 Premier League Season, Arsenal F.C. finished 5th in the premier league despite having the 3rd most wins. Furthermore, they also had the least amount of ties in the entirety of the premier league with a mere 3 draws out of 38 games. In this assignment, I aim to identify whether or not Arsenal had a more "aggressive" style of play compared to other teams in the Premeir League. This "aggressive" style may have lead to more volatile games in which it was less likely for a draw to be the outcome.

Variables of Interest

The variables of interest in this assignment are win percentage, full-time results, number of shots, and number of shots (all of which may be specific to being the home or away team). The variable of win percentage is not included in the data set but will be calculated in data wrangling process.

Data Wrangling

The first thing we will do is separate the data into three different "sets". The first containing all teams from the season in the object "all_teams", the second only including results from matches of teams who ended top ten in the Premier League in the object "top_ten", and the last one being the Arsenal specific set that contains only games from Arsenal in the object "Ars_dft".

```
all_teams <- df[4:18] #Subsets the data to columns that will be used for # analysis.

top_ten <- all_teams %>% #Data set involving only the games played by top 10 teams
  filter(HomeTeam == "Man City" | AwayTeam == "Man City" |
           HomeTeam == "Liverpool" | AwayTeam == "Liverpool" |
           HomeTeam == "Chelsea" | AwayTeam == "Chelsea" |
           HomeTeam == "Tottenham" | AwayTeam == "Tottenham" |
           HomeTeam == "Arsenal" | AwayTeam == "Arsenal"|
           HomeTeam == "Man United" | AwayTeam == "Man United"|
           HomeTeam == "West Ham" | AwayTeam == "West Ham"|
           HomeTeam == "Brighton" | AwayTeam == "Brighton"|
           HomeTeam == "Leicester City" | AwayTeam == "Leicester City"|
           HomeTeam == "Wolves" | AwayTeam == "Wolves"
         )


Ars_dft <- all_teams %>% #Subsets Arsenal specific data
  filter(HomeTeam == "Arsenal" | AwayTeam == "Arsenal") %>%
  mutate(H_v_A = ifelse(HomeTeam == 'Arsenal', 'H', 'A')) #Create new column
#that makes it easier to to identify whether Arsenal was Home or Away

Ars_dft <- Ars_dft %>%
  mutate(FTR = ifelse((HomeTeam == 'Arsenal' & FTR == 'H') |
                        (AwayTeam == 'Arsenal' & FTR == 'A'), 'W',
                      ifelse(FTR == 'D', 'D', 'L')))
#Changes Full Time Result column to whether or not Arsenal won, drew, or loss
```

```
head(Ars_dft) #Display new Arsenal data
```

```
##    HomeTeam  AwayTeam FTHG FTAG FTR HTHG HTAG HTR    Referee HS AS HST AST HF
## 1 Brentford   Arsenal    2    0   L    1    0   H   M Oliver  8 22   3   4 12
## 2   Arsenal   Chelsea    0    2   L    0    2   A  P Tierney  6 22   3   5 10
## 3  Man City   Arsenal    5    0   L    3    0   H M Atkinson 25  1  10   0  5
## 4   Arsenal   Norwich    1    0   W    0    0   D   M Oliver 30 10   7   1  9
## 5   Burnley   Arsenal    0    1   W    0    1   A   A Taylor 18 13   3   3  9
## 6   Arsenal Tottenham    3    1   W    3    0   H   C Pawson 12 10   7   4 12
##   AF H_v_A
## 1  8     A
## 2  4     H
## 3  7     A
## 4 11     H
## 5  8     A
## 6 13     H
```

Data Analysis

Investigation 1

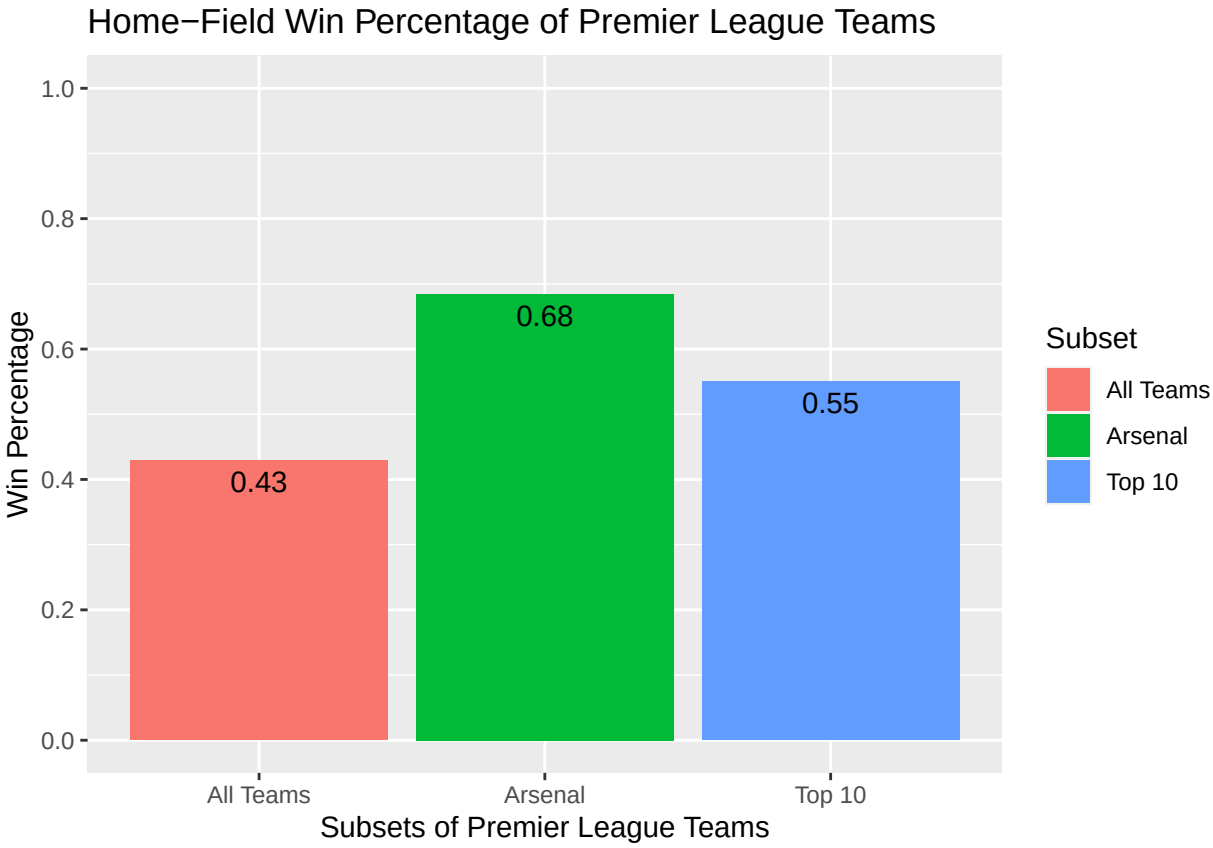
The first variable we will be investigating is whether or not Arsenal's home win record was high compared to other teams. This is sometimes referred to as a "home-field advantage" and could tell us if Arsenal had a stronger or weaker "home-field advantage".

```
combined_data <- bind_rows( #Creates new object combining data of pct home wins
Ars_dft %>% #Finds home win percentage of Arsenal
    filter(H_v_A == "H") %>%
    count(FTR) %>%
    pivot_wider(names_from = FTR, values_from = n) %>%
    mutate(win_pct = W/(W+L+D)),

all_teams %>% #Finds home win percentage from all teams
    count(FTR) %>%
    pivot_wider(names_from = FTR, values_from = n) %>%
    mutate(win_pct = H/(H+A+D)), #'H' is variable associated with Home wins

top_ten %>% #Finds home win percentage from top 10 teams
    filter(HomeTeam == "Man City" |
           HomeTeam == "Liverpool" |
           HomeTeam == "Chelsea" |
           HomeTeam == "Tottenham" |
           HomeTeam == "Arsenal" |
           HomeTeam == "Man United"|
             HomeTeam == "West Ham" |
             HomeTeam == "Brighton"|
             HomeTeam == "Leicester" |
             HomeTeam == "Wolves") %>%
    count(FTR) %>%
    pivot_wider(names_from = FTR, values_from = n) %>%
    mutate(win_pct = H/(H+A+D))
)
```

```r
rownames(combined_data) = c("Arsenal", "All Teams", "Top 10") #Labels new data set
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```r
#Plot to compare win percentages by the different subsets
combined_data %>%
  ggplot(aes(x = rownames(combined_data), y = win_pct, fill = rownames(combined_data))) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = round(win_pct, digits = 2)), vjust = 1.5) +
  scale_y_continuous(limits = c(0, 1), breaks = seq(0, 1, by = 0.2))+
  labs(title = "Home-Field Win Percentage of Premier League Teams",
       x = "Subsets of Premier League Teams",
       y = "Win Percentage",
       fill = "Subset")
```

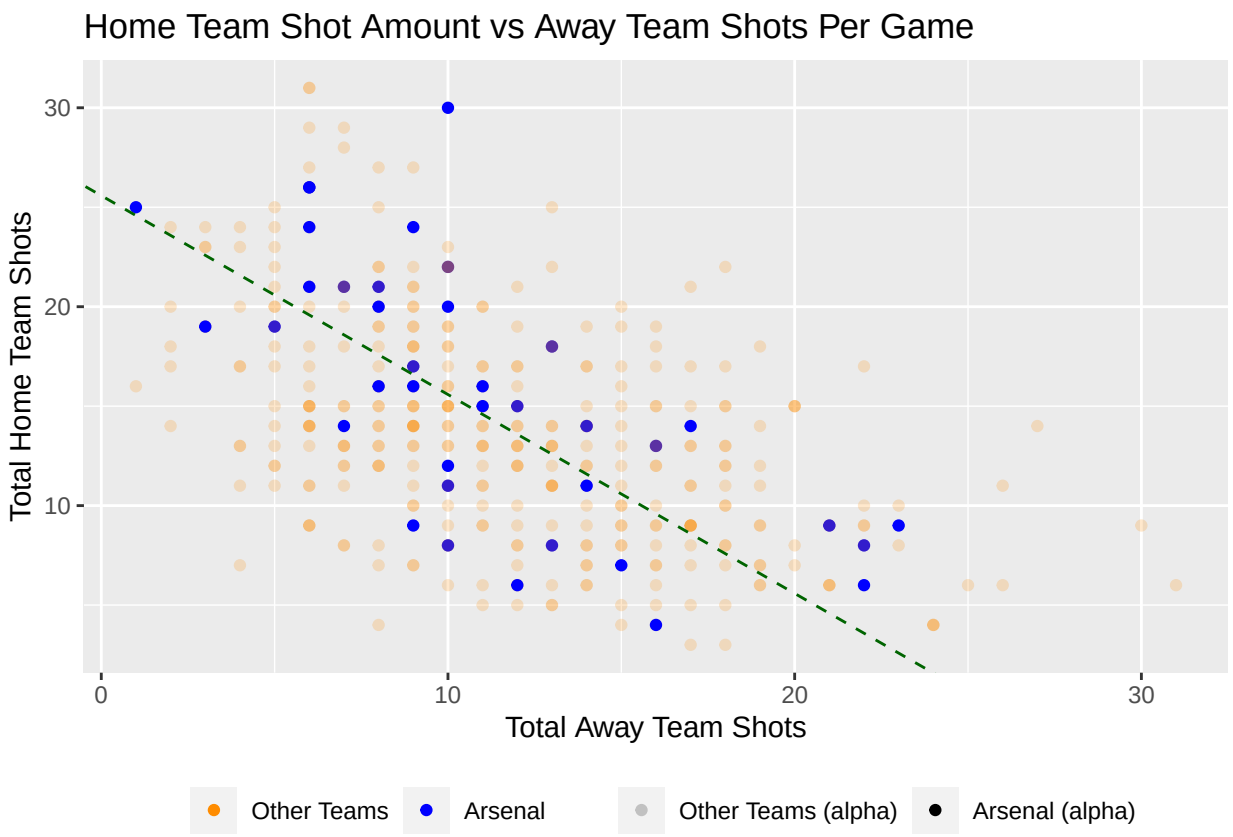


The barplot reveals to us that rounded the home win percentage of Arsenal is 68%, the top 10 teams is 55%, and the entire premier league is 43%. This is interesting since Arsenal finished 5th, it would make sense for their home win percentage to be closer to that of the entirety of the top 10 teams. These findings will be expanded on further in the discussion.

Investigation 2

Another factor we will be looking at is the number of shots that happened in Arsenal games. A possible theory for the few number of ties that occurred in Arsenal's season may have been because they had an aggressive play style; A play style that emphasizes offense over defense means we might see more total shots in Arsenal games compared to other teams in the Premier League.

```r
all_teams <- all_teams %>% #Add total shots column to data set
  mutate(total_shots = AS + HS)

all_teams %>% #Plot for Home Shots vs Away Shots
  ggplot(aes(x = AS,
             y = HS,
             color = (HomeTeam == 'Arsenal'| AwayTeam == 'Arsenal'),
             alpha = (HomeTeam == 'Arsenal'| AwayTeam == 'Arsenal'))) +
  scale_color_manual(values = c("TRUE" = "blue", "FALSE" = "dark orange"),
                     labels = c("TRUE" = "Arsenal",
                                "FALSE" = "Other Teams"))+
  scale_alpha_manual(values = c("TRUE" = 1, "FALSE" = 0.2),
                     labels = c("TRUE" = "Arsenal (alpha)",
                                "FALSE" = "Other Teams (alpha)")) +
  geom_point() +
  geom_abline(intercept = mean(all_teams$total_shots),
              slope = -1,
              linetype = "dashed", color = "dark green") +
  theme(legend.position = "bottom") +
  labs(title = str_wrap('Home Team Shot Amount vs Away Team Shots Per Game', width = 50),
  x = 'Total Away Team Shots',
  y = 'Total Home Team Shots',
  legend = NULL,
  color = element_blank(),
  alpha = element_blank())
```

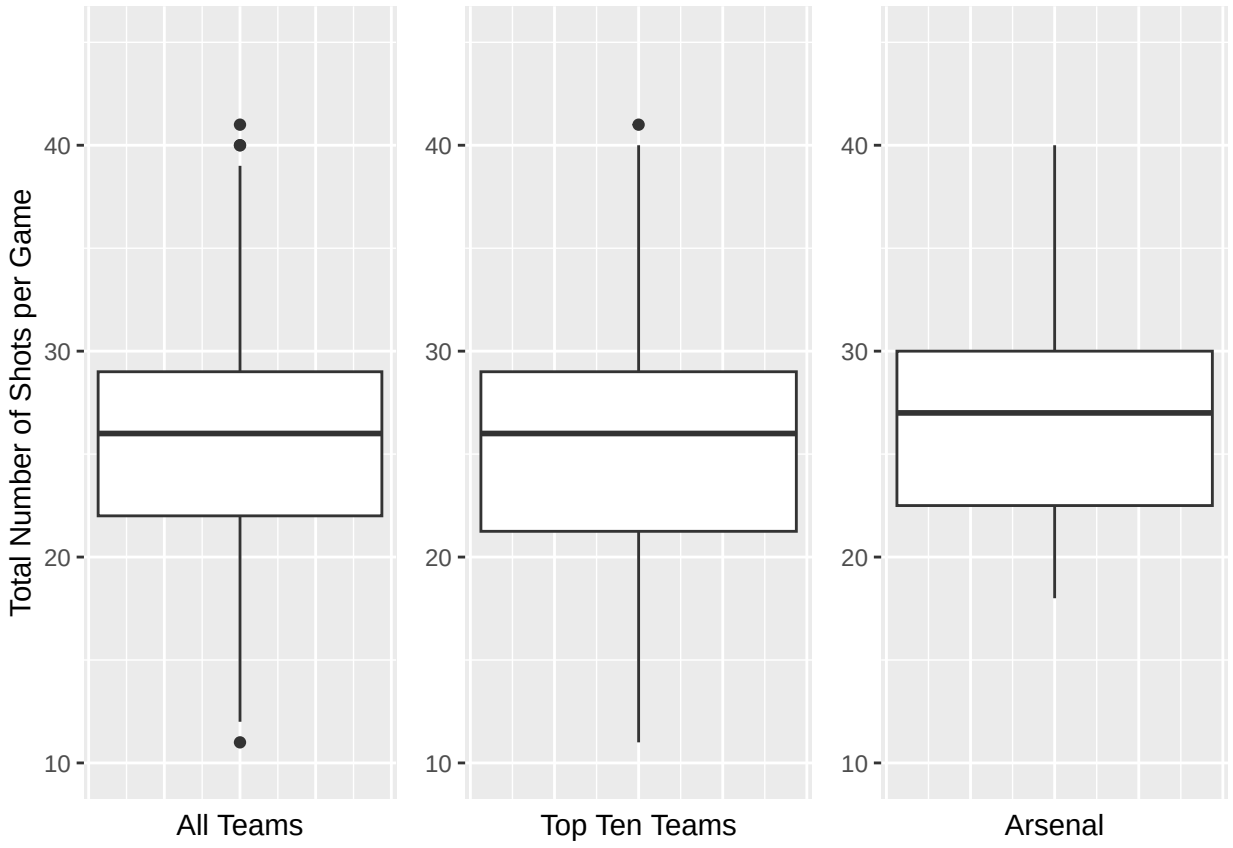


The graph shows the number of away shots vs home shots across all games in the Premier League with a

dashed green line indicating the average number of total shots per game. As we can see, a slight majority of the data points indicate when Arsenal are a playing team, are to the right of that dashed line. This means, that on average, Arsenal games follow this aforementioned "aggressive" play style as there are on average more total shots.

Boxplot Visualization

```
plot1 <- all_teams %>% #Boxplot of total shots for all teams
  mutate(total_shot = HS + AS) %>%
  ggplot(aes(y = total_shot)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(10, 45), breaks = seq(10, 45, by = 10)) +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
  labs(x = "All Teams", y = 'Total Number of Shots per Game')


plot2 <- top_ten %>% #Boxplot of total shots by the top ten teams
  mutate(total_shot = HS + AS) %>%
  ggplot(aes(y = total_shot)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(10, 45), breaks = seq(10, 45, by = 10)) +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
  labs(x = "Top Ten Teams", y = NULL)

plot3 <- Ars_dft %>% #Boxplot for total number of shots by Arsenal
  mutate(total_shot = HS + AS) %>%
  ggplot(aes(y = total_shot)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(10, 45), breaks = seq(10, 45, by = 10)) +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
  labs(x = "Arsenal", y = NULL)

final_plot <- grid.arrange(plot1, plot2, plot3, ncol = 3) #Combines all the boxplot graphs
```

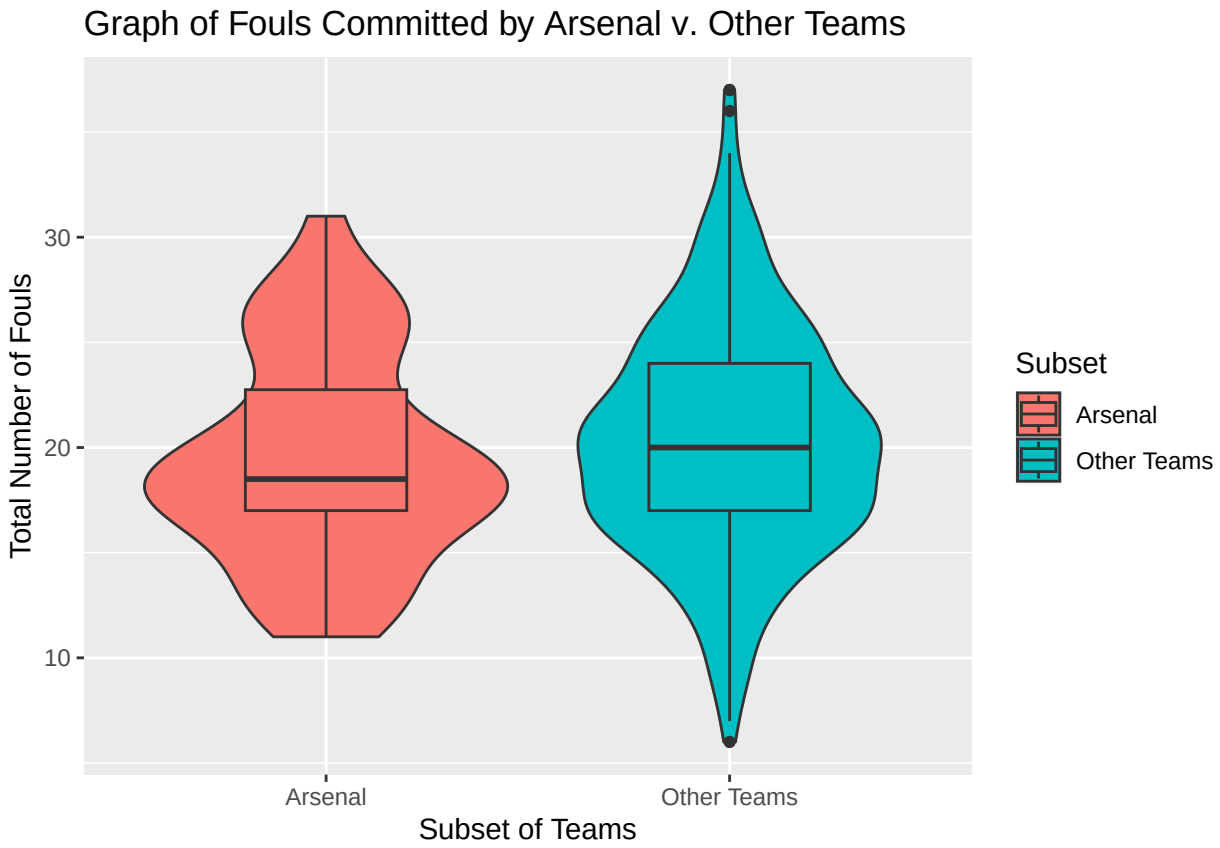


The boxplots reveal similar statistics across the 3 different subsets, with Arsenal having a slightly higher median average total shot per game compared to other two subsets. This once again might indicate the "aggressive" play style of Arsenal.

Investigation 3

The last investigation I will look at for this assignment is to see whether or not Arsenal played in games where more fouls were committed. Fouling is a way to stop counter-attacks from the opposing side. If Arsenal had a truly "aggressive" play style, this would mean they would have attacking players outnumber defensive players forcing the opposite team to foul, and in return, Arsenal would have to foul the opposition since they would be more susceptible to counter attacks.

```
all_teams %>% #graph of the violin plots
  mutate(Subset = ifelse(HomeTeam == 'Arsenal' | AwayTeam == 'Arsenal',
         'Arsenal', 'Other Teams')) %>%
  mutate(total_fouls = AF + HF) %>%
  ggplot(aes(x = Subset, y = total_fouls, fill = Subset))+
  geom_violin() +
  geom_boxplot(width = 0.4) +
  labs(title = "Graph of Fouls Committed by Arsenal v. Other Teams",
       x = "Subset of Teams",
       y = "Total Number of Fouls")
```



The violin plots reveal that Arsenal had in general a lower amount of fouls compared to the rest of the Premier League. This can be confirmed if we look at the boxplot which finds that the median number of total fouls in an Arsenal game is lower than the median number of total fouls in games from the rest of the Premier League.

Discussion

The data used in this assignment comes from the Premier League in which Arsenal finished 5th despite having the third most wins. However, our research question originates from the fact that Arsenal had the least amount of draws in the entirety of the Premier League with a mere 3. This made me interested in attempting to see if the reason behind these results from Arsenal were due to an "aggressive" play style: a play style characterized by overloading offense, leaving holes in the defense. I also looked at the home-field advantage of Arsenal to see if there was a great disparity between performance at home and performance at away stadiums for an explanation. Lastly, I had the data set split into 3 subsets (Arsenal, Top 10 teams, and entirety of Premier League) to better compare insights.

The first data visualization is a bar plot that takes the home-field win percentage of Arsenal compared to the other 3 subsets. The bar graph depicts that Arsenal had a 68% percent win rate at home compared to the league average 43% and top ten average of 55%. Clearly, Arsenal benefited heavily from playing at home, certainly moreso than other teams, and is certainly a statistic of interest considering it is 13% more than the top 10 team average win rate at home. This first investigation seems to support that Arsenal's efficiency in their home stadium greatly impacted their season final standings. In a future investigation it would be interesting to investigate their away win (or lose) percentage to see if there are any drastic differences there to account for the amount of losses sustained during Arsenal's season.

Investigation 2 was interested in seeing if Arsenal's play style was one that encouraged a greater amount of total shots in their matches. The dot plot showed that games in which Arsenal played that they were slightly more likely to be above the average total number of shots since a slight majority of the data points were to

the right of the dashed green line. Furthermore, the boxplots revealed a slightly higher median total of shots in Arsenal games compared to other two subsets. While this visualization doesn't seem the most significant as the magnitude of difference was marginal between Arsenal and other teams, it does provide a little insight of how offensively oriented these matches were.

The last investigation looked into see if there were a greater number of total fouls in Arsenal games due to their suspected "aggressive" play style. As noted before, an aggressive play style could be plausibly signaled by consistent fouling to stop counter-attacks by either team. After examining violin plots comparing the number of fouls committed in Arsenal games compared to the rest of the league, there appeared to be no significant differences. It could be possible there is no correlation between fouls committed and play style of teams since refereeing is not necessarily standardized across games. Furthermore working from a sample size of 38 games might not be enough to prove a statistically significant relationship or correlation between the two variables.

Ultimately, this investigation provides perhaps a couple of insights as to why Arsenal performed the way they did in the 2021-2022 Premier League season. Arsenal's home season win percentage is certainly a statistic of a top-tier club and the number of shots may indicate some aspects of how they play. However this investigation may be limited by not doing enough to research why they loss as many games as they did. Trying to define a vague term like "aggressive" in relation to play styles also doesn't have a standardized quantitative measure as it was referred to more qualitatively throughout the investigation. Furthermore, subsetting the data the way I did may have not been the most efficient way to subset data for statistical insights.

References

The original dataset was found on kaggle.com with the link:

https://www.kaggle.com/datasets/saife245/english-premier-league/data.

ChatGPT was used to help find binding row functionality for grouping the boxplots and bar plot data visualizations, as well as helping create the line on the home shot vs away shot graph. R-help resources online were also used to help with formatting of ggplots as well.