---
title: "The Referee Report: A Data Science Analysis of Premier League Referees"
author: "Written by Alex Marzban, Alicja Mahr, Arya Antara, Lance Yoon, Mark Kim, Ryan Lee"
output: github_document

Introduction

Refereeing decisions have always been an important part of football since they are responsible for enforcing the rules and ensuring fair play. However, there are instances where the decisions made by referees can impact the outcome of the game. Over the years, there has been a lot of debate about the impact that referees can have on the outcome of a match. Especially with the recent World Cup, this debate has been heightened, with some high-profile incidents leading to controversy and calls for changes to the way that referees are selected, trained, and evaluated.

One way in which referees can affect the outcome of a football game is through their interpretation of the rules. Referees have a certain amount of discretion when it comes to enforcing the laws of the game, and their interpretations can sometimes be subjective. For example, a referee may choose to overlook a minor foul committed by a player, while another referee may consider it to be a more serious offense and award a free kick or even a penalty kick. These decisions can have a significant impact on the game since a free kick or penalty kick can change the momentum and ultimately the outcome of the match.

Another way in which referees can affect the outcome of the game is through their mistakes. Referees are human, and they can make mistakes in judgment or miss critical incidents during the game. However, these mistakes can also have a significant impact on the match’s outcome. In some cases, a referee’s mistake can result in a goal being awarded to the wrong team or a crucial penalty being overlooked, which can lead to an unfair advantage for one team over the other.

There are a lot of ways in which referees can have a significant impact on the outcome of the game. Therefore, our group wanted to focus on the 2014-2019 seasons of the English Premier League and analyze the impact the referees had for those two seasons. In order to do this, we collected and integrated five datasets containing various information about the matches, including the match results, player statistics, and referee performance.We then compared the frequency of each referee giving out yellow and red cards, fouls, free kicks, etc. to determine how the referees have an impact on the outcome of premier league matches. Our main objective for this project was to determine referee biases and the amount of impact they put on the outcome of games.

Questions

1. Is there a difference between the number of fouls, yellow cards, and red cards booked against the Big 6 relative to other teams?
1. Which referees display bias towards certain teams?
1. Which teams are most likely to have fouls, yellow cards, and red cards booked against them?

Exploration
```{r Data and Libraries, echo=FALSE, message=FALSE}
library(tidyverse)
library(GGally)
library(skimr)
library(gridExtra)
library(cowplot)
library(gt)

epl_data <- read.csv("data/epl_2014_2019.csv")
referee_counts <- table(epl_data$Referee)
qualified_referees <- names(referee_counts[referee_counts >= 20])

# Filter the epl_data data frame to only include games refereed by qualified referees
filtered_epl_data <- epl_data[epl_data$Referee %in% qualified_referees, ]
epl_data <- filtered_epl_data
```

~~~~
```

```{r Correlation, echo=FALSE, message=FALSE, fig.align='center'}
epl_data %>%
  select(c(FTHG, FTAG, HTHG, HTAG)) %>%
  ggpairs(cardinality_threshold = 50, progress = FALSE)+
  labs(caption="Figure 1")+
  theme(plot.caption = element_text(hjust = 0.5))
```
```{r Red and Yellow Cards per Match, echo=FALSE, message=FALSE, fig.align='center'}
ref_cards <- epl_data %>%
  group_by(Referee) %>%
  transmute(YellowCard = HY + AY, RedCard = HR + AR) %>%
  summarise(yellow = sum(YellowCard) / n(), red = sum(RedCard) / n())


# summarize(ref_cards, games_reffed = n())
ggplot(ref_cards, aes(Referee, yellow)) +
  geom_bar(stat = "identity", position = "dodge",fill="gold1") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(title="Yellow Card Ratio vs Referee",y="Yellow Card Ratio per Match")+
  labs(caption="Figure 2")+
  theme(plot.caption = element_text(hjust = 0.5))
ggplot(ref_cards, aes(Referee, red)) +
  geom_bar(stat = "identity", position = "dodge",fill="red") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(title="Red Card Ratio vs Referee",y="Red Card Ratio per Match")+
  labs(caption="Figure 3")+
  theme(plot.caption = element_text(hjust = 0.5))
```
```{r Manchester United Heat Map, echo=FALSE, message=FALSE, fig.align='center'}
#  create a new dataframe with only the relevant columns
card_data <- epl_data %>%
  select(HomeTeam, AwayTeam, Referee, HY, AY)

# separate the data into two groups: cards given to the home team and cards given to the
away team
home_cards <- card_data %>%
  transmute(team = HomeTeam, opp_team = AwayTeam, cards = HY, Referee) %>%
  mutate(location = "home")
away_cards <- card_data %>%
  transmute(team = AwayTeam, opp_team = HomeTeam, cards = AY, Referee) %>%
  mutate(location = "away")

# combine the two groups into a single dataframe
all_cards <- rbind(home_cards, away_cards)

# create a new column for the total number of cards given to each team by each referee
all_cards <- all_cards %>%
  group_by(team, Referee) %>%
  summarize(total_cards = sum(cards)) %>%
  filter(!is.na(total_cards)) %>%
  ungroup()

# create a new column for the total number of cards given to each team by all referees
team_totals <- all_cards %>%
  group_by(team) %>%
  summarize(total_cards = sum(total_cards)) %>%
  filter(!is.na(total_cards)) %>%
  ungroup()

# calculate the percentage of cards given to the home team by each referee
card_percentages <- all_cards %>%
~~~~

````
  left_join(team_totals, by = "team") %>%
  mutate(card_percentage = total_cards.x / total_cards.y) %>%
  select(-total_cards.x, -total_cards.y)

# create a heat map for each home team
# teams <- unique(card_data$HomeTeam)
x <- "Man United"
team_cards <- card_percentages %>%
  filter(team == x)

ggplot(team_cards, aes(x = reorder(Referee, -card_percentage), y = card_percentage)) +
  geom_point( size=3, color='gold1') +
  labs(x = "Referee", y = "Ratio of Cards to Home Team", title = paste("Home Team:", x)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(caption="Figure 4")+
  theme(plot.caption = element_text(hjust = 0.5))



```
```{r Yellow Card Breakdown, echo=FALSE, message=FALSE, fig.align='center'}
# load data
card_data <- epl_data %>%
  select(HomeTeam, AwayTeam, Referee, HY, AY)

# create separate dataframes for home and away cards
home_cards <- card_data %>%
  mutate(location = "home") %>%
  select(HomeTeam, Referee, HY, location) %>%
  rename(team = HomeTeam, cards = HY)

away_cards <- card_data %>%
  mutate(location = "away") %>%
  select(AwayTeam, Referee, AY, location) %>%
  rename(team = AwayTeam, cards = AY)

# combine the dataframes
all_cards <- rbind(home_cards, away_cards)

# calculate total cards for each team and each referee
team_totals <- all_cards %>%
  group_by(Referee, location) %>%
  summarize(total_cards = sum(cards) / n()) %>%
  filter(!is.na(total_cards)) %>%
  ungroup()

# calculate percentage of cards for each team and each referee
card_percentages <- team_totals %>%
  group_by(Referee, location) %>%
  summarize(card_percentage = mean(total_cards)) %>%
  ungroup()

# plot yellow card percentages for home vs. away
ggplot(team_totals %>% group_by(Referee, location), aes(x = Referee, y = total_cards, fill = location)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("home" = "blue", "away" = "red")) +
  labs(x = "Referee", y = "Yellow Card Per Match", fill = "", title="Yellow Card Per Match vs Referee") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(caption="Figure 5")+
  theme(plot.caption = element_text(hjust = 0.5))
````

````
```

```{r Red Card Breakdown, echo=FALSE, message=FALSE, fig.align='center'}
# load data
card_data <- epl_data %>%
  select(HomeTeam, AwayTeam, Referee, HR, AR)

# create separate dataframes for home and away cards
home_cards <- card_data %>%
  mutate(location = "home") %>%
  select(HomeTeam, Referee, HR, location) %>%
  rename(team = HomeTeam, cards = HR)

away_cards <- card_data %>%
  mutate(location = "away") %>%
  select(AwayTeam, Referee, AR, location) %>%
  rename(team = AwayTeam, cards = AR)

# combine the dataframes
all_cards <- rbind(home_cards, away_cards)

# calculate total cards for each team and each referee
team_totals <- all_cards %>%
  group_by(Referee, location) %>%
  summarize(total_cards = sum(cards) / n()) %>%
  filter(!is.na(total_cards)) %>%
  ungroup()

# calculate percentage of cards for each team and each referee
card_percentages <- team_totals %>%
  group_by(Referee, location) %>%
  summarize(card_percentage = mean(total_cards)) %>%
  ungroup()

# plot yellow card percentages for home vs. away
ggplot(team_totals %>% group_by(Referee, location), aes(x = Referee, y = total_cards, fill
= location)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("home" = "blue", "away" = "red")) +
  labs(x = "Referee", y = "Red Card Per Match", fill = "",title="Red Card Per Match vs
Referee") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(caption="Figure 6")+
  theme(plot.caption = element_text(hjust = 0.5))

```
```{r Foul Breakdown, echo=FALSE, message=FALSE, fig.align='center'}
data <- epl_data %>%
  filter(!is.na(AY), !is.na(HY), !is.na(HF), !is.na(AF)) %>%
  group_by(Referee) %>%
  summarize(AF = sum(AF) / n(), HF = sum(HF) / n(), HY = sum(HY) / n(), AY = sum(AY) / n())
%>%
  pivot_longer(cols = c(AF, HF), names_to = "Location", values_to = "Fouls")
ggplot(data, aes(x = Referee, y = Fouls, fill = Location)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Referee", y = "Fouls Per Match", title = "Fouls vs Referee") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_fill_discrete(labels = c("Home Fouls", "Away Fouls"))+
  labs(caption="Figure 7")+
  theme(plot.caption = element_text(hjust = 0.5))
```
````
````

```{r Yellow Cards per Foul,echo=FALSE, message=FALSE, warning=FALSE, fig.align='center'}
# Step 3: Preprocess the data (updated)
epl_data_processed <- epl_data %>%
  group_by(Referee) %>%
  filter(!is.na(AY), !is.na(HY), !is.na(HF), !is.na(AF)) %>%
  summarise(Yellow = AY + HY, Fouls = HF + AF, Ratio = Yellow / Fouls)

# Step 4: Create the boxplot
ggplot(epl_data_processed, aes(x = Referee, y = Ratio, group = Referee)) +
  geom_boxplot() +
  labs(x = "Referee", y = "Yellow Cards per Foul", fill = "", title = "Yellow Card Ratio vs
Referee") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(caption="Figure 8")+
  theme(plot.caption = element_text(hjust = 0.5))
```

Results

Question 1: Is there a difference between the number of fouls, yellow cards, and red cards booked against the Big 6 relative to other teams?

```{r q1-stats, warning=FALSE, echo=FALSE, fig.align='center'}
# Count the number of fouls, yellow cards, and red cards booked against the Big 6 and Bad 6
teams
big_6 <- c("Arsenal", "Chelsea", "Liverpool", "Man City", "Man United", "Tottenham")
bad_6 <- c("Middlesbrough", "QPR", "Fulham", "Cardiff", "Wolves", "Huddersfield")

# Create a new dataframe with only the relevant columns
card_data <- epl_data %>%
  select(HomeTeam, AwayTeam, HF, AF, HY, AY, HR, AR)

# Separate the data into two groups: cards given to the home team and cards given to the
away team
home_cards <- card_data %>% transmute(team=HomeTeam, HF, HY, HR) %>%
  group_by(team) %>%
  summarize(home_fouls = sum(HF), home_yellow_cards = sum(HY), home_red_cards = sum(HR))

away_cards <- card_data %>% transmute(team=AwayTeam, AF, AY, AR) %>%
  group_by(team) %>%
  summarize(away_fouls = sum(AF), away_yellow_cards = sum(AY), away_red_cards = sum(AR))

# Combine the two groups into a single dataframe
all_cards <- left_join(home_cards, away_cards, by = "team")

# Calculate the total number of fouls, yellow cards, and red cards per game for each team
all_cards <- all_cards %>%
  mutate(total_fouls = home_fouls + away_fouls,
         total_yellow_cards = home_yellow_cards + away_yellow_cards,
         total_red_cards = home_red_cards + away_red_cards) %>%
  select(-home_fouls, -away_fouls, -home_yellow_cards, -away_yellow_cards, -home_red_cards,
-away_red_cards)

# Calculate the number of games played by each team
home_games_played <- epl_data %>%
  group_by(HomeTeam) %>%
  summarize(h_games_played = n()) %>%
  rename(team = HomeTeam)

away_games_played <- epl_data %>%
  group_by(AwayTeam) %>%
  summarize(a_games_played = n()) %>%
  rename(team = AwayTeam)
```

```
# Combine the two groups into a single dataframe
games_played <- left_join(home_games_played, away_games_played, by = "team")

# Calculate the total number of games played by each team
games_played <- games_played %>%
  mutate(total_games_played = h_games_played + a_games_played) %>%
  select(-h_games_played, -a_games_played)

# Calculate the average number of fouls, yellow cards, and red cards per game for each team
all_cards <- left_join(all_cards, games_played, by = "team") %>%
  mutate(avg_fouls = total_fouls / total_games_played,
         avg_yellow_cards = total_yellow_cards / total_games_played,
         avg_red_cards = total_red_cards / total_games_played) %>%
  select(-total_fouls, -total_yellow_cards, -total_red_cards, -total_games_played)

# Separate the data into two groups: cards given to the Big 6 and cards given to the Bad 6
big_6_cards <- all_cards %>%
  filter(team %in% big_6)

bad_6_cards <- all_cards %>%
  filter(team %in% bad_6)

CI <- function(data, alpha) {
  x_bar <- mean(data)
  s <- sd(data)
  n <- length(data)
  q <- qnorm(1 - alpha / 2)
  lower <- x_bar - q * s / sqrt(n)
  upper <- x_bar + q * s / sqrt(n)
  return(c(lower, upper))
}

# Determine if there is a significant difference between the number of fouls, yellow cards,
and red cards booked against the Big 6 and Bad 6 teams
big_6_fouls <- mean(big_6_cards$avg_fouls)
bad_6_fouls <- mean(bad_6_cards$avg_fouls)
big_6_yellow_cards <- mean(big_6_cards$avg_yellow_cards)
bad_6_yellow_cards <- mean(bad_6_cards$avg_yellow_cards)
big_6_red_cards <- mean(big_6_cards$avg_red_cards)
bad_6_red_cards <- mean(bad_6_cards$avg_red_cards)

# Calculate the z-score for the difference in the average number of fouls, yellow cards,
and red cards per game for the Big 6 and Bad 6
z_fouls <- (big_6_fouls - bad_6_fouls) / sqrt(var(big_6_cards$avg_fouls) /
length(big_6_cards$avg_fouls) + var(bad_6_cards$avg_fouls) / length(bad_6_cards$avg_fouls))
z_yellow_cards <- (big_6_yellow_cards - bad_6_yellow_cards) /
sqrt(var(big_6_cards$avg_yellow_cards) / length(big_6_cards$avg_yellow_cards) +
var(bad_6_cards$avg_yellow_cards) / length(bad_6_cards$avg_yellow_cards))
z_red_cards <- (big_6_red_cards - bad_6_red_cards) / sqrt(var(big_6_cards$avg_red_cards) /
length(big_6_cards$avg_red_cards) + var(bad_6_cards$avg_red_cards) /
length(bad_6_cards$avg_red_cards))

# Calculate the p-value for the difference in the average number of fouls, yellow cards,
and red cards per game for the Big 6 and Bad 6
p_fouls <- 2 * pnorm(-abs(z_fouls))
p_yellow_cards <- 2 * pnorm(-abs(z_yellow_cards))
p_red_cards <- 2 * pnorm(-abs(z_red_cards))

# Calculate the 95% confidence interval for the difference in the average number of fouls,
yellow cards, and red cards per game for the Big 6 and Bad 6
big_6_bad_6_fouls_CI <- CI(big_6_cards$avg_fouls - bad_6_cards$avg_fouls, 0.05)
big_6_bad_6_yellow_cards_CI <- CI(big_6_cards$avg_yellow_cards -
bad_6_cards$avg_yellow_cards, 0.05)
big_6_bad_6_red_cards_CI <- CI(big_6_cards$avg_red_cards - bad_6_cards$avg_red_cards, 0.05)
```

````
summary_df <- data.frame(
  metric = c("Fouls", "Yellow Cards", "Red Cards"),
  big_6_mean = c(big_6_fouls, big_6_yellow_cards, big_6_red_cards),
  bad_6_mean = c(bad_6_fouls, bad_6_yellow_cards, bad_6_red_cards),
  p_value = c(p_fouls, p_yellow_cards, p_red_cards),
  lower_CI = c(big_6_bad_6_fouls_CI[1], big_6_bad_6_yellow_cards_CI[1],
big_6_bad_6_red_cards_CI[1]),
  upper_CI = c(big_6_bad_6_fouls_CI[2], big_6_bad_6_yellow_cards_CI[2],
big_6_bad_6_red_cards_CI[2])
)
summary_table <- gt(summary_df) %>%
  tab_header(
    title = "Summary Statistics",
    subtitle = "Figure 9"
  ) %>%
  cols_align(align = "left") # Set column alignment to "left"

summary_table

```

```{r big 6, echo=FALSE, message=FALSE, warning=FALSE, fig.align='center'}


# Define the big 6 Premier League teams
big_6 <- c("Arsenal", "Chelsea", "Liverpool", "Man City", "Man United", "Tottenham")

# Filter the dataset for the big 6 teams and calculate the overall median of fouls
epl_data_big_6 <- epl_data %>%
  filter(HomeTeam %in% big_6) %>%
  mutate(Team = as.factor(HomeTeam), Ratio = HY / HF)
epl_dataa <- epl_data %>% mutate(Ratio = HY / HF)
overall_median_fouls <- median(epl_dataa$Ratio, na.rm = TRUE)

# Create box plots for each of the big 6 teams and add a horizontal line for the overall
median of fouls
ggplot(epl_data_big_6, aes(x = Team, y = Ratio, fill = Team)) +
  geom_boxplot() +
  geom_hline(yintercept = overall_median_fouls, linetype = "dashed", color = "blue", size =
1) +
  labs(x = "Home Team", y = "Ratio of Yellow Cards to Fouls", title = "Box Plots of Ratio
of Yellow Cards to Fouls vs Home Team for the Big 6 Premier League Teams", fill = "Team") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(caption="Figure 10")+
  theme(plot.caption = element_text(hjust = 0.5))

```

## Question 2: Which referees display bias towards certain teams?

```{r q2 graphs yellow, echo=FALSE, message=FALSE, fig.align='center'}

# Which referees display bias towards certain teams?
# Create a new dataframe with only the relevant columns
card_data <- epl_data %>%
  select(HomeTeam, AwayTeam, Referee, HY, AY)

# Separate the data into two groups: cards given to the home team and cards given to the
away team
home_cards <- card_data %>% transmute(team=HomeTeam, Referee, HY) %>%
  group_by(Referee, team) %>%
  summarize(cards = sum(HY), totalx = n())

away_cards <- card_data %>% transmute(team=AwayTeam, Referee, AY) %>%
````

````
  group_by(Referee, team) %>%
  summarize(cards = sum(AY), totaly = n())

# Combine the two groups into a single dataframe
all_cards <- left_join(home_cards, away_cards, by = c("Referee", "team"))

# Calculate the average number of yellow cards per game for each Referee
all_cards <- all_cards %>%
  mutate(total_games = ifelse(is.na(totalx), 0,totalx)+ ifelse(is.na(totaly), 0, totaly), #
calculate total games by referee
        total_cards = (ifelse(is.na(cards.x),0,cards.x) + ifelse(is.na(cards.y),0,cards.y))
/ total_games) %>% # calculate average cards per game by referee
  select(-totalx, -totaly, -cards.x, -cards.y) # remove unnecessary columns
# Get list of unique referees
referees <- unique(card_percentages$Referee)
referees <- c("A Marriner", "L Mason", "M Atkinson", "R East")

# Create an empty list to store the plots
yellow_card_plots <- list()

for (x in referees) {
  team_cards <- all_cards %>%
    filter(Referee == x)

  plot <- ggplot(team_cards, aes(x = reorder(team, -total_cards), y = total_cards)) +
    geom_bar(stat = "identity", fill="gold1") +
    labs(x = "Team", y = "Mean Yellows per Game", title = paste("Referee:", x)) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))

  # Append the plot to the list
  yellow_card_plots[[x]] <- plot
}

# Arrange the plots in a 2x2 grid
combined_plot <- plot_grid(plotlist = yellow_card_plots, ncol = 2)

# Add the label to the bottom center of the grid
final_plot <- ggdraw(combined_plot) +
  draw_label("Figure 11", x = 0.5, y = 0.025, hjust = 0.5, vjust = 0.5, size = 8)

final_plot


```

```{r q2 graphs red, warning=FALSE, echo=FALSE, message=FALSE, fig.align='center'}

# Which referees display bias towards certain teams?
# Create a new dataframe with only the relevant columns
card_data <- epl_data %>%
  select(HomeTeam, AwayTeam, Referee, HR, AR)

# Separate the data into two groups: cards given to the home team and cards given to the
away team
home_cards <- card_data %>% transmute(team=HomeTeam, Referee, HR) %>%
  group_by(Referee, team) %>%
  summarize(cards = sum(HR), totalx = n())

away_cards <- card_data %>% transmute(team=AwayTeam, Referee, AR) %>%
  group_by(Referee, team) %>%
  summarize(cards = sum(AR), totaly = n())

# Combine the two groups into a single dataframe
all_cards <- left_join(home_cards, away_cards, by = c("Referee", "team"))
````

```
# Calculate the average number of yellow cards per game for each Referee
all_cards <- all_cards %>%
  mutate(total_games = ifelse(is.na(totalx), 0,totalx)+ ifelse(is.na(totaly), 0, totaly), #
calculate total games by referee
        total_cards = (ifelse(is.na(cards.x),0,cards.x) + ifelse(is.na(cards.y),0,cards.y))
/ total_games) %>% # calculate average cards per game by referee
  select(-totalx, -totaly, -cards.x, -cards.y) # remove unnecessary columns
# Get list of unique referees
referees <- unique(card_percentages$Referee)
referees <- c("A Marriner", "A Taylor", "R East", "M Atkinson")

# Create a graph for each referee
# Create an empty list to store the plots
red_card_plots <- list()

for (x in referees) {
  team_cards <- all_cards %>%
    filter(Referee == x)

  plot <- ggplot(team_cards, aes(x = reorder(team, -total_cards), y = total_cards)) +
    geom_bar(stat = "identity", fill="red") +
    labs(x = "Team", y = "Mean Reds per Game", title = paste("Referee:", x)) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))

  # Append the plot to the list
  red_card_plots[[x]] <- plot
}

# Arrange the plots in a 2x2 grid
combined_plot <- plot_grid(plotlist = red_card_plots, ncol = 2)

# Add the label to the bottom center of the grid
final_plot <- ggdraw(combined_plot) +
  draw_label("Figure 12", x = 0.5, y = 0.025, hjust = 0.5, vjust = 0.5, size = 8)

final_plot

```

````
```{r q2 graphs fouls, echo=FALSE, warning=FALSE, message=FALSE, fig.align='center'}

# Which referees display bias towards certain teams?
# Create a new dataframe with only the relevant columns
card_data <- epl_data %>%
  select(HomeTeam, AwayTeam, Referee, HF, AF)

# Separate the data into two groups: cards given to the home team and cards given to the
away team
home_cards <- card_data %>% transmute(team=HomeTeam, Referee, HF) %>%
  group_by(Referee, team) %>%
  summarize(cards = sum(HF), totalx = n())

away_cards <- card_data %>% transmute(team=AwayTeam, Referee, AF) %>%
  group_by(Referee, team) %>%
  summarize(cards = sum(AF), totaly = n())

# Combine the two groups into a single dataframe
all_cards <- left_join(home_cards, away_cards, by = c("Referee", "team"))

# Calculate the average number of yellow cards per game for each Referee
all_cards <- all_cards %>%
  mutate(total_games = ifelse(is.na(totalx), 0,totalx)+ ifelse(is.na(totaly), 0, totaly), #
calculate total games by referee
````
````

```
        total_cards = (ifelse(is.na(cards.x),0,cards.x) + ifelse(is.na(cards.y),0,cards.y))
/ total_games) %>% # calculate average cards per game by referee
  select(-totalx, -totaly, -cards.x, -cards.y) # remove unnecessary columns
# Get list of unique referees
referees <- unique(card_percentages$Referee)
referees <- c("A Marriner", "A Taylor", "R East", "M Atkinson")

# Create a graph for each referee
plots <- list()

for (x in referees) {
  team_cards <- all_cards %>%
    filter(Referee == x)

  plot <- ggplot(team_cards, aes(x = reorder(team, -total_cards), y = total_cards)) +
    geom_bar(stat = "identity", fill="blue") +
    labs(x = "Team", y = "Mean Fouls per Game", title = paste("Referee:", x)) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))

  # Append the plot to the list
  plots[[x]] <- plot
}
# Arrange the plots in a 2x2 grid
combined_plot <- plot_grid(plotlist = plots, ncol = 2)

# Add the label to the bottom center of the grid
final_plot <- ggdraw(combined_plot) +
  draw_label("Figure 13", x = 0.5, y = 0.025, hjust = 0.5, vjust = 0.5, size = 8)

final_plot
```

Question 3: Which teams are most likely to have fouls, yellow cards, and red cards booked against them?

```{r q3 graphs, echo=FALSE, message=FALSE, fig.align='center'}

# add yellow_per_game and red_per_game to the data frame
epl_data <- epl_data %>%
  group_by(HomeTeam) %>%
  mutate(
    fouls_per_game = HF / n(),
    yellow_per_game = HY / n(),
    red_per_game = HR / n()
  ) %>%
  ungroup()

# Create a graph of the average fouls per match for each team, sorting the x axis by the
average number of fouls
ggplot(epl_data, aes(x = reorder(HomeTeam, -fouls_per_game), y = fouls_per_game)) +
  geom_bar(stat = "identity", position = "stack",fill="blue") +
  labs(x = "Home Team", y = "Fouls per Match", title = "Fouls by Home Team") +
  scale_fill_manual(values = c("blue"), labels = c("Fouls")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(caption="Figure 14")+
  theme(plot.caption = element_text(hjust = 0.5))

# Create a graph of the average yellow cards per match for each team
ggplot(epl_data, aes(x = reorder(HomeTeam, -yellow_per_game), y = yellow_per_game)) +
  geom_bar(stat = "identity", position = "stack",fill="gold1") +
  labs(x = "Home Team", y = "Yellow Cards per Match", title = "Yellow Cards by Home Team")
+
  scale_fill_manual(values = c("gold1"), labels = c("Yellow Cards")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
```

```
  labs(caption="Figure 15")+
  theme(plot.caption = element_text(hjust = 0.5))

# Create a graph of the average red cards per match for each team
ggplot(epl_data, aes(x = reorder(HomeTeam, -red_per_game), y = red_per_game)) +
  geom_bar(stat = "identity", position = "stack",fill="red") +
  labs(x = "Home Team", y = "Red Cards per Match", title = "Red Cards by Home Team") +
  scale_fill_manual(values = c("red"), labels = c("Red Cards")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(caption="Figure 16")+
  theme(plot.caption = element_text(hjust = 0.5))
```

Discussion

Exploration

In our analysis of the English Premier League (EPL), we focused on two separate areas: the relationship between half-time and full-time goal counts and the number of yellow cards issued by each referee to the home and away teams. Initially, we created a catalog of qualified referees who had officiated in at least 20 games over the five seasons we considered. We then filtered the 'epl_data' data frame, retaining only the games officiated by these qualified referees, and performed exploratory analysis on the relationship among the half-time and full-time goal counts for both home and away teams.

Our analysis showed that, apart from the correlation between half-time home goals (HTHG) vs full-time home goals (FTHG) and half-time away goals (HTAG) vs full-time away goals (FTAG), there was no significant correlation among other variables. The correlation coefficient for the former comparison was 0.7, and the scatter plot demonstrated that when a team scored more goals in the first half, they were more likely to maintain their lead and win the game.

Moving on to the second area of focus, we computed the number of yellow cards given to both teams by each referee and created a heat map to show the ratio of cards given to the home team vs the total number of cards in a game by each referee. We split the data into two groups - 'home_cards' and 'away_cards' - and created a single data frame, 'all_cards,' that included the total number of cards given to each team by each referee. We created a new variable to represent the total number of cards given to each team by all referees and calculated the percentage of cards given to the home team by each referee. The heat map created for Manchester United revealed a significant difference between the referees.

Our analysis suggests that referees' decisions can impact the outcome of games, with a team that is consistently given more yellow cards than their opponents facing a potential disadvantage. However, note that our analysis only considered yellow cards and did not include other factors that could affect the game's outcome, such as red cards, penalties, or injuries. Furthermore, there may be lurking variables that could affect the results. Despite these limitations, the heat map gives an accurate picture of which referees tend to give more yellow cards to home teams, providing valuable information for future research.

Furthermore, we also looked at data about yellow and red cards given out by referees in the English Premier League (EPL). We chose the information we needed and made two separate lists, one for yellow cards given to the home team and one for yellow cards given to the away team. We combined these two lists to make a single list of all the yellow cards given out. We also did the same thing for red cards. We used this data to make a graph showing which referees gave out the most and least yellow and red cards to home and away teams.

Our analysis shows that Phil Dowd gave the most yellow cards overall, and Mike Dean gave the most red cards to away teams. In addition, the referees who gave out the least number of cards were Paul Tierney for red cards to away teams, and Mike Jones for red cards to home teams.

We made a box plot to compare how often referees give out yellow cards in the English Premier League. This helps us see if there are any patterns in how referees make decisions. We looked at data for each referee, which included information about how many yellow cards and fouls happened in each game. We calculated a ratio of yellow cards to fouls for each referee and used this to create the box plot. We found that most referees gave out yellow cards less often than average, which might mean that players were committing fewer fouls or less serious fouls. The referee with the highest percentage of yellow cards was Mike Dean, and the referee with the lowest was Paul Tierney. We also noticed one referee, Neil Swarbick, had an extreme outlier value. This could mean he was officiating games with more fouls.

To check if referees have a bias towards home teams in calling fouls, a study selected 20 referees and compared the number of fouls they gave to home and away teams. Most referees gave around 10 fouls, with P. Dowd giving the most and A. Marrier giving the least. The analysis suggested that referees tend to give more fouls to the home team. 15 referees gave more fouls to the home team than the away team, indicating a possible bias. However, it's important to keep in mind that other factors such as player skills and luck also play a role in game outcomes. It's also important to investigate if this trend exists in other sports and regions.

Results

Question 1: Is there a difference between the number of fouls, yellow cards, and red cards booked against the Big 6 relative to other teams?

In order to answer the first research question, we separated the data into two groups, cards given to the home team and cards given to the away team. The total number of fouls, yellow cards, and red cards per game for each team was calculated. The number of games played by each team was calculated, and the average number of fouls, yellow cards, and red cards per game for each team was calculated. Finally, the data was separated into two groups, cards given to the Big 6 and cards given to the Bad 6.

The findings are as follows. The average number of fouls per game for the Big 6 was 10.4659. The z-score for the difference in the average number of fouls per game for the Big 6 and Bad 6 was -1.2814, which corresponds to a p-value of 0.2001. This indicates that the difference between the number of fouls per game for the Big 6 and Bad 6 teams in the EPL is not statistically significant at a significance level of 0.05.

The average number of yellow cards per game for the Big 6 was 1.6097, and for the Bad 6 was 1.8119. The z-score for the difference in the average number of yellow cards per game for the Big 6 and Bad 6 was -2.2133, which corresponds to a p-value of 0.0269. This indicates that there is a significant difference between the number of yellow cards per game for the Big 6 and Bad 6 teams in the EPL.

The average number of red cards per game for the Big 6 was 0.0583, and for the Bad 6 was 0.0533. The z-score for the difference in the average number of red cards per game for the Big 6 and Bad 6 was 0.3697, which corresponds to a p-value of 0.7116. This indicates that there is no significant difference between the number of red cards per game for the Big 6 and Bad 6 teams in the EPL.

The results show that there is a significant difference between the number of fouls and red cards per game for the Big 6 and Bad 6 teams in the EPL, but there is not a significant difference between the number of yellow cards per game for the Big 6 and Bad teams. However, since we cannot replicate this process with the other lower teams in the league, we cannot say this is significant.

Question 2: Which referees display bias towards certain teams?

These series of bar charts show the average number of yellow cards per game for each referee and the teams they've reffed for, sorted by greatest to least. To perform this, we first created a new dataframe, card_data, from epl_data, which contained only our desired columns (HomeTeam, AwayTeam, Referee, HY, AY). Next, we divided each referee's total number of cards given by the total number of games each of them reffed for. A ggplot was then created, with the X-axis representing each team and the Y-axis representing the average yellow cards per game.

One notable bar chart include R East's, who gave on average of over 4 yellow cards per game to Wolverhampton Wanderers (Wolves) and Middlesbrough FC. In comparison, the 3rd highest was just over 2 yellow cards per game given to Aston Villa. Another interesting bar chart came from M Atkinson, who gave approximately 3 yellow cards per game to Queens Park Rangers, a noticable amount more than the second most he gave to Bumley at less than 2.5 yellow cards. On the other side of the spectrum, referee L Mason gave Swansea an average of less than .5 yellow cards per games, which is much less than the 3 yellow cards per game they gave to Cardiff. However, when looking at the other referee's bar graphs, in which most teams range between 1-2 yellow cards per game for all referees, based on this statistic alone, we cannot definitively conclude that all or even certain referees have a strong biased towards certain teams.

Similarly with the yellow cards, these graphs depict the average number of red cards per game for each referee and the teams they've reffed for, sorted by greatest to least. The process of creating them is the same, with the exception of selecting HR and AR (home and away red cards) instead of HY and AY. The X-axis represents the each team and the Y-axis represents the average red cards per game.

If we look at referee R East again, we observe that he gives roughly .5 red cards per game to Hull and .3 to Brighton and Huddersfield. In contrast with the yellow cards, they did not give any red cards to the Wolves or Middlesborough. One notable referee was M Atkinson, who gave Queens Park Rangers (QPR) the most yellow and red cards per game (~3 and ~.25, respectively) compared to the other teams they've reffed for.

These bar graphs show the average number of fouls each referee calls per game for each team they've reffed for. The dataframe was created using the same method as the yellow and red cards, except selecting for home fouls (HF) and away fouls (AF). The X-axis represents the teams, and the Y-axis represents the average number of fouls per game. It's important to note that the number of red and yellow cards don't necessarily add up to the total number of fouls per game, as players more often than not are simply given verbal warnings.

Looking at referee R East again, they called the most fouls on Middlesbrough (~15 per game), which is one of the two teams they gave the most yellow cards to on average. However, this didn't necessarily correlate with the number of red cards they handed, as he never gave Middlesbrough a single red card over the 5 seasons. According to M Atkinson's foul data, he called the second most fouls on QPR at over 10, which could possibly indicate a negative bias they have towards the team, when also considering the number of yellow and red cards they give to that team. When looking at the other referees however, there doesn't seem to be a clear correlation between the average number of yellow and red cards and the number of fouls called for each team per game, as their results vary greatly.

Question 3: Which teams are most likely to have fouls, yellow cards, and red cards booked against them?

This analysis uses three different bar charts to represent the average yellow cards, red cards, and fouls per match for every home teams. The code starts with adding three columns to the `epl_data` data frame: `fouls_per_game`, `yellow_per_game`, and `red_per_game`. These columns are created by grouping the data by HomeTeam and calculating the average number of fouls, yellow, and red cards per match. After adding these three new variables, three bar charts are created with the x-axis representing each home teams and the y-axis representing the average number of fouls, yellow, and red cards per match.

The three charts show that different teams have varying levels of fouls, yellow, and red cards, which could impact the outcome of the game. We can observe from the graphs that some teams have a higher number of fouls and cards per game than other teams. However, we can observe that the average number of fouls and the average number of yellow cards booked by the home teams are generally similar to the rest of other teams but we can observe that Watford has the highest number average fouls (~12.5) and yellow cards (~2.25) from the recorded dataset. The last bar graph that shows the average number of red cards per match for each team and we can observe that Huddersfield, Aston Villa, and Hull had the highest average number of red cards of around 0.125 red cards per game. However, unlike the previous two bar graphs, this bar graph shows that the number of red cards received per match is very different depending on the home teams. For example, although Tottenham commits an average of ~11 fouls per match and an average of ~1.4 yellow cards per match, they are only booked for red cards around 0.0125 times per match. However, we can notice that some of the Bad 6 teams have a low percentage of receiving red cards per game. Therefore, these three charts do not constitute towards outcome of matches.

This result of the analysis suggests that the average number of fouls, yellow, and red cards given per match does not necessarily correlate with the outcome of premier league matches. Some teams commit more fouls, receive more yellow or red cards than other teams. Although the analysis does not allow us to confidently conclude that referees significantly impact the outcome of games, we can conclude that having high percentage of receiving fouls, yellow, or red cards have a minuscule impact on the outcome of matches, as there are countless tactical fouls committed in a match.

Limitations

There are several possible lurking variables that could impact the study's findings on whether referees exhibit a home advantage bias in calling fouls in sports. For example, weather conditions can affect the playing conditions, which may lead to more fouls being committed. Similarly, rivalry matches may be more intense and emotionally charged, potentially causing players to commit more fouls and referees to make more disciplinary decisions. Therefore, these variables should be considered and accounted for in the analysis to ensure that the results are not skewed.

It's also important to note that the introduction of the Visual Assistant Referee (VAR) in the EPL after 2019 may have affected the results. Before VAR, referees had to make decisions subjectively, based on their interpretation of the events on the pitch. However, with VAR, there is a video review system that can help referees make more accurate decisions. This change in the refereeing process could have had an impact on the frequency and distribution of fouls and cards, and thus, should be considered when interpreting the study's results.

Furthermore, the study's findings may be limited by the sample size for certain teams. For example, teams that are frequently relegated or promoted to or from the EPL may not have enough data to draw meaningful conclusions about their disciplinary patterns. This could introduce bias into the analysis, and the results may not be representative of the broader population of EPL teams. Therefore, it's important to ensure that the sample size is large enough and that all teams are adequately represented to minimize potential biases and improve the generalizability of the findings.