

Project 1

Shubham Khalkho

2024-06-07

Intro and Problem

This report will be analyzing a Premier League Dataset. The English Premier League is the highest and most popular football league in the world.

- This data set contains all the games from 1993 to present.
- All the dates are included with the timings and Home and Away
- All the goals scored by both teams and what were the results
- The problem I will be addressing will be how to predict using set pieces to tell that which teams will win. (Basically a winning calculator)

Loading the data

```
# Loading necessary libraries
library(dplyr)

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##   filter, lag
## The following objects are masked from 'package:base':
##   intersect, setdiff, setequal, union
library(readr)
library(ggplot2)
library(plotly)

##
## Attaching package: 'plotly'
## The following object is masked from 'package:ggplot2':
##   last_plot
## The following object is masked from 'package:stats':
##   filter
## The following object is masked from 'package:graphics':
##
```

```

##      layout
library(tidyr)

#Loading the Data
data <- read_csv("PremierLeague.csv")

## Rows: 11780 Columns: 39

## -- Column specification -----
## Delimiter: ","
## chr (8): Season, Date, Time, HomeTeam, AwayTeam, FullTimeResult, HalfTimeRe...
## dbl (31): FullTimeHomeTeamGoals, FullTimeAwayTeamGoals, HalfTimeHomeTeamGoal...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

colnames(data)

## [1] "Season"                  "Date"                    "Time"
## [4] "HomeTeam"                 "AwayTeam"                 "FullTimeHomeTeamGoals"
## [7] "FullTimeAwayTeamGoals"   "FullTimeResult"          "HalfTimeHomeTeamGoals"
## [10] "HalfTimeAwayTeamGoals"   "HalfTimeResult"          "Referee"
## [13] "HomeTeamShots"           "AwayTeamShots"           "HomeTeamShotsOnTarget"
## [16] "AwayTeamShotsOnTarget"   "HomeTeamCorners"         "AwayTeamCorners"
## [19] "HomeTeamFouls"            "AwayTeamFouls"            "HomeTeamYellowCards"
## [22] "AwayTeamYellowCards"     "HomeTeamRedCards"        "AwayTeamReadCards"
## [25] "B365HomeTeam"             "B365Draw"                "B365AwayTeam"
## [28] "B365Over2.5Goals"       "B365Under2.5Goals"       "MarketMaxHomeTeam"
## [31] "MarketMaxDraw"           "MarketMaxAwayTeam"        "MarketAvgHomeTeam"
## [34] "MarketAvgDraw"           "MarketAvgAwayTeam"        "MarketMaxOver2.5Goals"
## [37] "MarketMaxUnder2.5Goals"  "MarketAvgOver2.5Goals"   "MarketAvgUnder2.5Goals"

```

Data Cleaning, Wrangling and Munging

The data has been put into R from the csv file but some of the data in the file is not in the format that I want. For example the Data types for Date and its format is incorrect. So we will be formatting the data into Date Month and Year. Apart from it a lot of values are filled with null. Instead of Keeping it that way I will be filling it up with the median values of the Column for better data and easier plotting.

```

#Formatting Data.
data$date <- as.Date(data$date, format = "%d/%m/%Y")
#Filling empty values with 0
data <- data %>%
  mutate(across(everything(), ~ifelse(is.na(.), 0, .)))

```

Distribution of Full-Time Home Team Goals

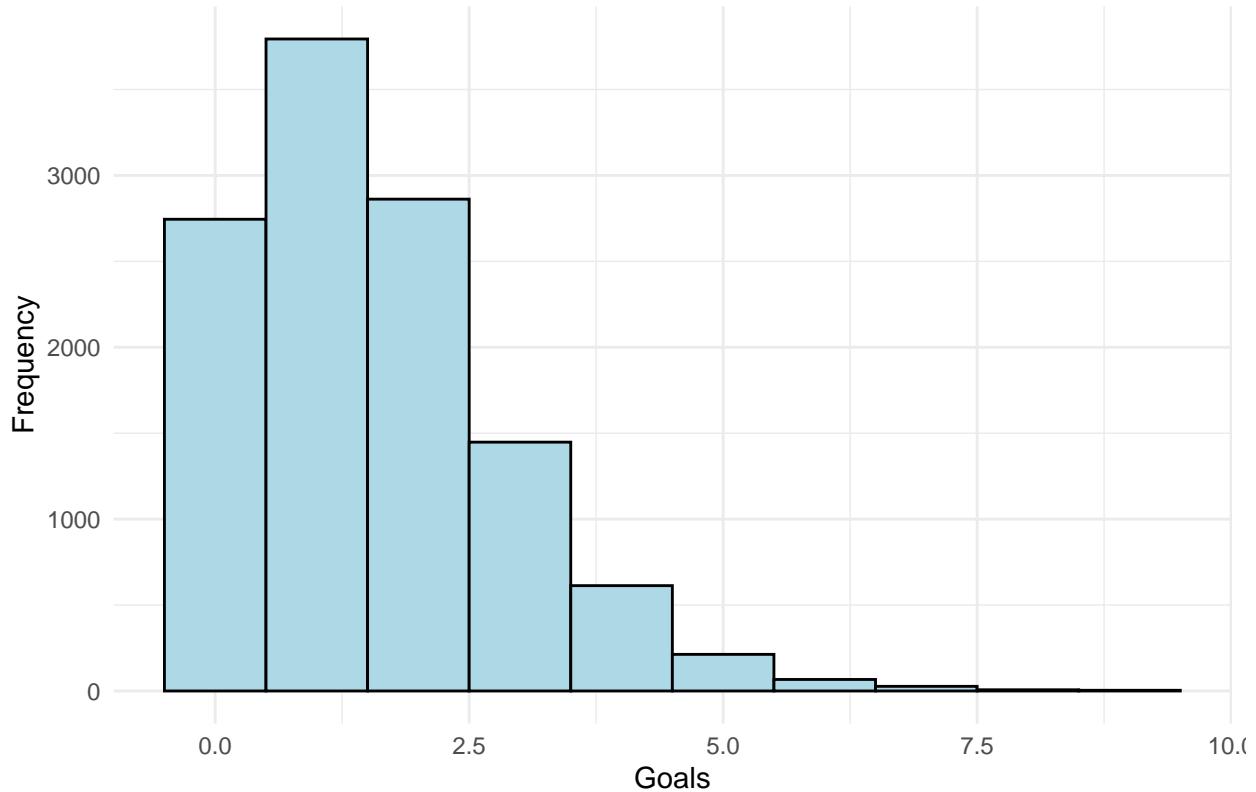
```

## Distribution of Full-Time Home Team Goals

ggplot(data, aes(x = FullTimeHomeTeamGoals)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  labs(title = "Distribution of Full-Time Home Team Goals", x = "Goals", y = "Frequency") +
  theme_minimal()

```

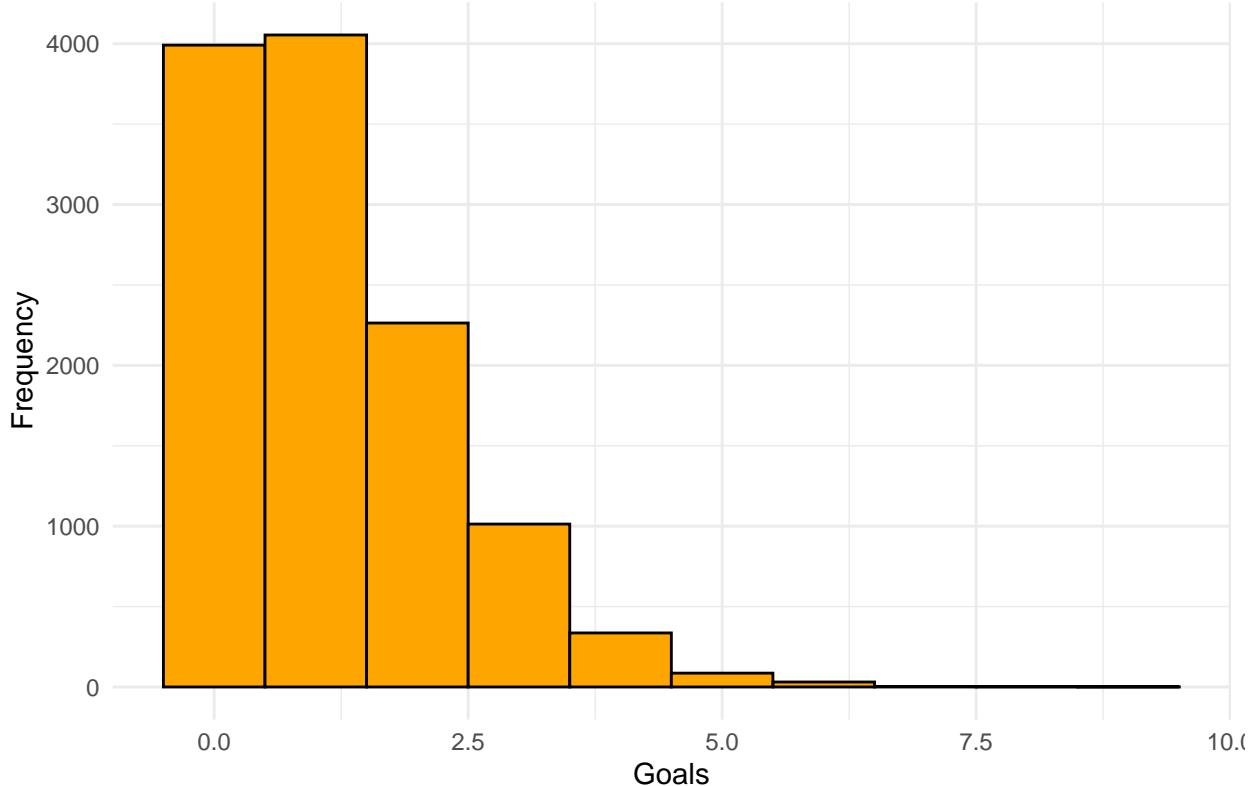
Distribution of Full-Time Home Team Goals



Distribution of Full-Time Away Team Goals

```
ggplot(data, aes(x = FullTimeAwayTeamGoals)) +  
  geom_histogram(binwidth = 1, fill = "orange", color = "black") +  
  labs(title = "Distribution of Full-Time Away Team Goals", x = "Goals", y = "Frequency") +  
  theme_minimal()
```

Distribution of Full-Time Away Team Goals



Here Graphs 1 and 2 show on an average how many goals are scored in a full time of a football match. We see a pattern where there is a difference in the number of goals scored from certain ranges. The Away teams tend to score more on the lower side of the scale whereas the home teams tend to score more on the upper side of the scale under 2.5. The average shows us that they score almost the same amount of goals but in different ranges.

Home Team Goals vs. Away Team Goals with Linear Regression

```
# Fit the linear regression model
model <- lm(FullTimeAwayTeamGoals ~ FullTimeHomeTeamGoals, data = data)

# Summary of the linear regression model
summary(model)

## 
## Call:
## lm(formula = FullTimeAwayTeamGoals ~ FullTimeHomeTeamGoals, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.2396 -1.1313 -0.1855  0.8145  7.7604 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.239647  0.016219 76.430 < 2e-16 ***
## FullTimeHomeTeamGoals -0.054149  0.008042 -6.733 1.74e-11 ***
```

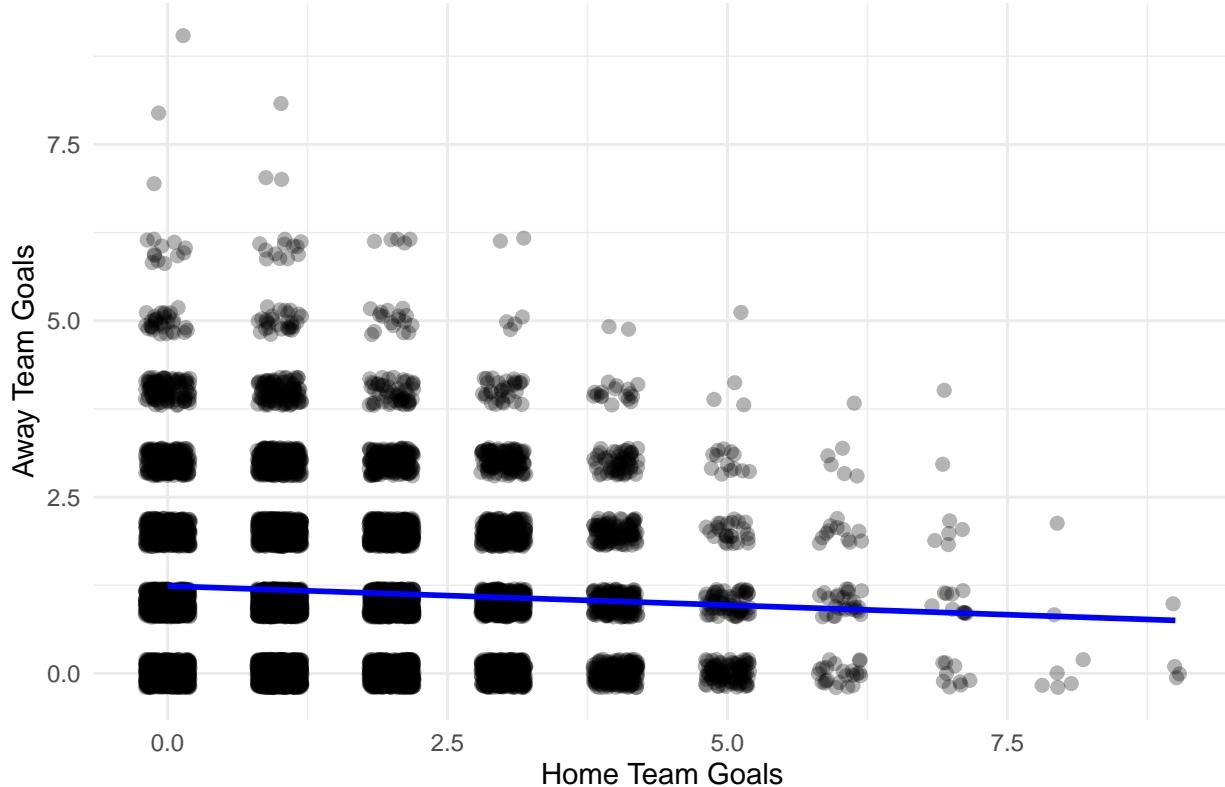
```

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.144 on 11778 degrees of freedom
## Multiple R-squared:  0.003834, Adjusted R-squared:  0.00375
## F-statistic: 45.34 on 1 and 11778 DF, p-value: 1.737e-11
# Plotting the relationship with the regression line, with jitter
ggplot(data, aes(x = FullTimeHomeTeamGoals, y = FullTimeAwayTeamGoals)) +
  geom_jitter(alpha = 0.3, size = 2, width = 0.2, height = 0.2) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Home Team Goals vs. Away Team Goals with Linear Regression",
       x = "Home Team Goals", y = "Away Team Goals") +
  theme_minimal()

## `geom_smooth()` using formula = 'y ~ x'

```

Home Team Goals vs. Away Team Goals with Linear Regression



In this plot we see that all the goals scored by both the teams are almost the same which indicates that a lot of the matches ended in a draw instead of a winner. We see that the number of goals are concentrated in the area of this 4 X 4 cube which is around 2.5 goals. After that it starts to fade as the it is rare for both the teams to score that big of a number of goals in the game

In the next plots we will be looking at how many times the home team won and how many times the away team won. The question we raised in the starting of the project was to see whether we can quantify the betting odds on the game. We will see how many times this was successful and how many times this was unsuccessful. By using all the data we will be quantifying and making a model which solves this everlasting problem of how the odds on this game change and what should be done to win most of the times.

Combined Plot of Average Goals Scored and Wins

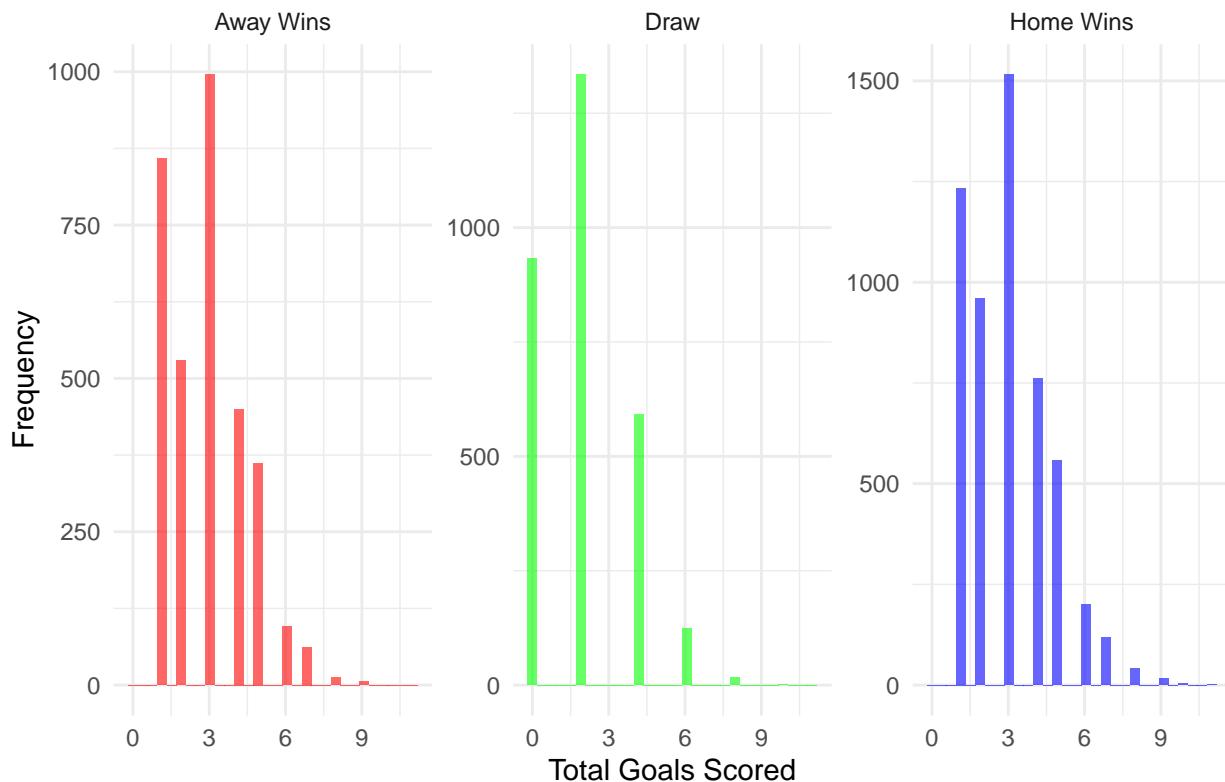
```
data <- data %>%
  mutate(
    HomeWin = ifelse(FullTimeHomeTeamGoals > FullTimeAwayTeamGoals, 1, 0),
    AwayWin = ifelse(FullTimeAwayTeamGoals > FullTimeHomeTeamGoals, 1, 0),
    Draw = ifelse(FullTimeHomeTeamGoals == FullTimeAwayTeamGoals, 1, 0)
  )

# Prepare data for plotting
plot_data <- data %>%
  mutate(Result = case_when(
    HomeWin == 1 ~ "Home Wins",
    AwayWin == 1 ~ "Away Wins",
    Draw == 1 ~ "Draw"
  ))

# Create the faceted histograms
ggplot_plot <- ggplot(plot_data, aes(x = FullTimeHomeTeamGoals + FullTimeAwayTeamGoals, fill = Result))
  geom_histogram(bins = 30, alpha = 0.6) +
  labs(title = "Distribution of Goals Scored in Home Wins, Away Wins, and Draws",
       x = "Total Goals Scored",
       y = "Frequency") +
  scale_fill_manual(values = c("Home Wins" = "blue", "Away Wins" = "red", "Draw" = "green")) +
  theme_minimal() +
  theme(legend.position = "none", legend.title = element_blank()) +
  facet_wrap(~ Result, scales = "free_y")

# Print the ggplot
print(ggplot_plot)
```

Distribution of Goals Scored in Home Wins, Away Wins, and Draws



```

# Calculate the total number of matches played
total_matches <- nrow(data)

# Print the total number of matches played
total_matches

## [1] 11780

# Calculate the outcomes
outcome_counts <- data %>%
  summarise(
    Home_Wins = sum(FullTimeHomeTeamGoals > FullTimeAwayTeamGoals),
    Away_Wins = sum(FullTimeAwayTeamGoals > FullTimeHomeTeamGoals),
    Draws = sum(FullTimeHomeTeamGoals == FullTimeAwayTeamGoals)
  )

# Print the results
outcome_counts

## # A tibble: 1 x 3
##   Home_Wins Away_Wins Draws
##       <int>     <int> <int>
## 1      5410      3370  3000

```

We see that the total number of games played is 11780 where almost 50% of the matches won were by the home teams. Showing that out of that big of a number the away teams won only 3370 times and the game between both the teams was drawn approximately 3000 times. From the graphs and data it is far more likely for the home teams to win than the away team in the Premier League. So the best option that we have is to rely on the home teams to win. It is a rarer occurrence for the away team to win and the data supports

it so much so that it is almost as likely for the game to draw as well.

Now that we have set up a proper description of how many goals are scored, how many times does the home team win, how many times the away team wins and how many times does the game draws. It is common knowledge that when a game is won the team which wins scores more goals than the team which loses so to see which team will be the winner this is not the best parameter to see that. This is more of a conclusion as to how many times does a team win or draw and how many goals are scored on an average. The best way to see that what team will be winning will be talked in the next part.

Calculating on the basis of set pieces

Now we will be seeing how do the statistics affect the number of goals scored and how this accounts to winning. This will include things like corners, yellow cards, red cards, total shots taken by a team, total shots on target taken by a team. We will see how each of these factors contribute to winning the game. If it does matter we will see what criteria matters the most so that we can get a clear vision of what parameter to consider when we want to see what team will be winning.

The most important statistic we have is the total shots taken and shots on target. It is well known that if you shoot more and shoot more accurately you tend to score more goals. But is it truly the case. Lets find out using a few graphs.

Graph on Red Cards given to both the teams and who wins that game.

```
# Check for the actual column names
data <- data %>%
  rename(AwayTeamRedCards = AwayTeamReadCards)

# Filter and transform data for red card scenarios
red_card_data <- data %>%
  filter(HomeTeamRedCards > 0 | AwayTeamRedCards > 0) %>%
  mutate(
    HomeTeamRedCards = ifelse(is.na(HomeTeamRedCards), 0, HomeTeamRedCards),
    AwayTeamRedCards = ifelse(is.na(AwayTeamRedCards), 0, AwayTeamRedCards),
    RedCardScenario = case_when(
      HomeTeamRedCards > 0 & AwayTeamRedCards > 0 ~ "Both Teams",
      HomeTeamRedCards > 0 & AwayTeamRedCards == 0 ~ "Home Team",
      HomeTeamRedCards == 0 & AwayTeamRedCards > 0 ~ "Away Team"
    ),
    FullTimeResult = case_when(
      FullTimeHomeTeamGoals > FullTimeAwayTeamGoals ~ "Home Win",
      FullTimeHomeTeamGoals < FullTimeAwayTeamGoals ~ "Away Win",
      FullTimeHomeTeamGoals == FullTimeAwayTeamGoals ~ "Draw"
    )
  )

# Summarize the data
summary_red_card_data <- red_card_data %>%
  group_by(RedCardScenario, FullTimeResult) %>%
  summarise(Count = n(), .groups = 'drop')

# Create the plot
red_card_plot <- ggplot(summary_red_card_data, aes(x = RedCardScenario, y = Count, fill = FullTimeResult))
```

```

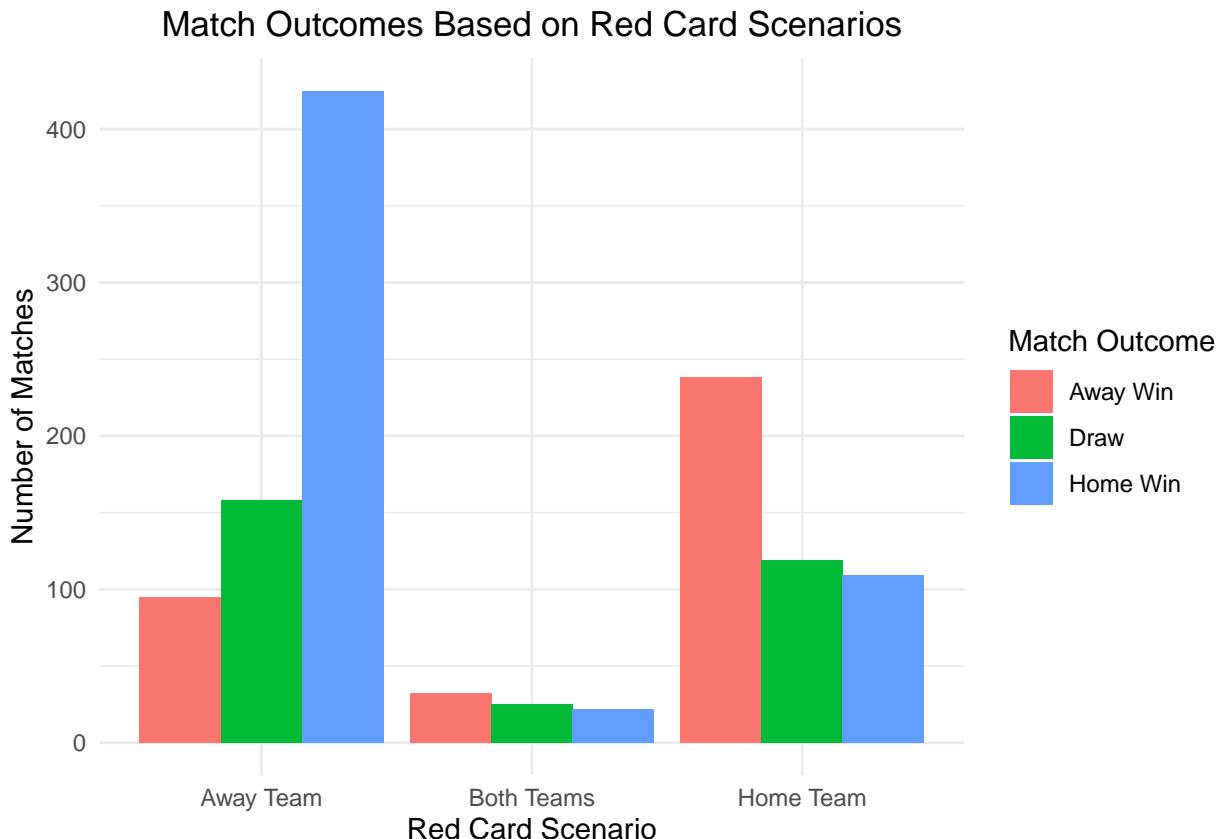
geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Match Outcomes Based on Red Card Scenarios",
       x = "Red Card Scenario",
       y = "Number of Matches",
       fill = "Match Outcome") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

# Convert ggplot to plotly
red_card_plotly <- ggplotly(red_card_plot)

# Print the plot
print(red_card_plotly)

# Generate the plot using ggplot
print(red_card_plot)

```



```

# Filter and transform data for red card scenarios
red_card_data <- data %>%
  filter(HomeTeamRedCards > 0 | AwayTeamRedCards > 0) %>%
  mutate(
    HomeTeamRedCards = ifelse(is.na(HomeTeamRedCards), 0, HomeTeamRedCards),
    AwayTeamRedCards = ifelse(is.na(AwayTeamRedCards), 0, AwayTeamRedCards),
    RedCardScenario = case_when(
      HomeTeamRedCards > 0 & AwayTeamRedCards > 0 ~ "Both Teams",
      HomeTeamRedCards > 0 & AwayTeamRedCards == 0 ~ "Home Team",
      HomeTeamRedCards == 0 & AwayTeamRedCards > 0 ~ "Away Team"
    )
  )

```

```

)
)

# Summary of red card counts for each scenario
red_card_summary <- red_card_data %>%
  group_by(RedCardScenario) %>%
  summarise(
    TotalRedCards = sum(HomeTeamRedCards) + sum(AwayTeamRedCards)
  )

# Print the summary
print(red_card_summary)

## # A tibble: 3 x 2
##   RedCardScenario TotalRedCards
##   <chr>           <dbl>
## 1 Away Team        703
## 2 Both Teams       168
## 3 Home Team        485

```

The pattern here is that whenever either of the team gets a red card the opposition wins mostly. In the first part of of the graph where we see the number of red cards given to away teams, we see that the home team wins most of the times. The away team gets a red card a total of 703 times where the home team wins about 400 + times. They draw about 150 times and the away teams win approximately 100 times.

Whenever both the teams get a red card the away team wins more than the home teams. The teams draws more than the home wins and the away team wins more than the drawing. But the values are very close by. Showing that there is no significant difference.

The pattern for the third graph is that whenever the home team gets a read card the away team wins most of the time. The total numer of times the home teams gets a red card is about 485 times. The Away team wins approximately 250 times. The match ends in a draw around 125 times and th home teams wins less than 125 times.

The pattern here is that the away teams tend to get more red cards in a game and lose more in comparision to the home teams getting a red card. We see that the number of red cards heavily influence the flow of the game and affect the winning tendency of a team. We see that that Red Cards given out to away teams show us some sort of bias when we see the game in its entirety. It could be to win the favor of the public or it could be the away teams just fouling the home team more.

Parametric for Shots taken and Shots on Target

Another parametric we can use to finalize which wins is the number of Shots Taken and How many of those are on Target. If you want to score a goal you need to shoot on the target which is the goalpost. This might be the most effective way to see which team will win in a given game. After this I will be plotting a Graph which takes into consideration, all of the data we considered and quantify what attributes to winning the most and what is the trend for a winner in the game.

Graphs on Shots Taken and Shots on Target and How it Affects the Winner

```

# Calculate average shots and shots on target for each match
match_summary <- data %>%
  group_by(FullTimeHomeTeamGoals, FullTimeAwayTeamGoals) %>%

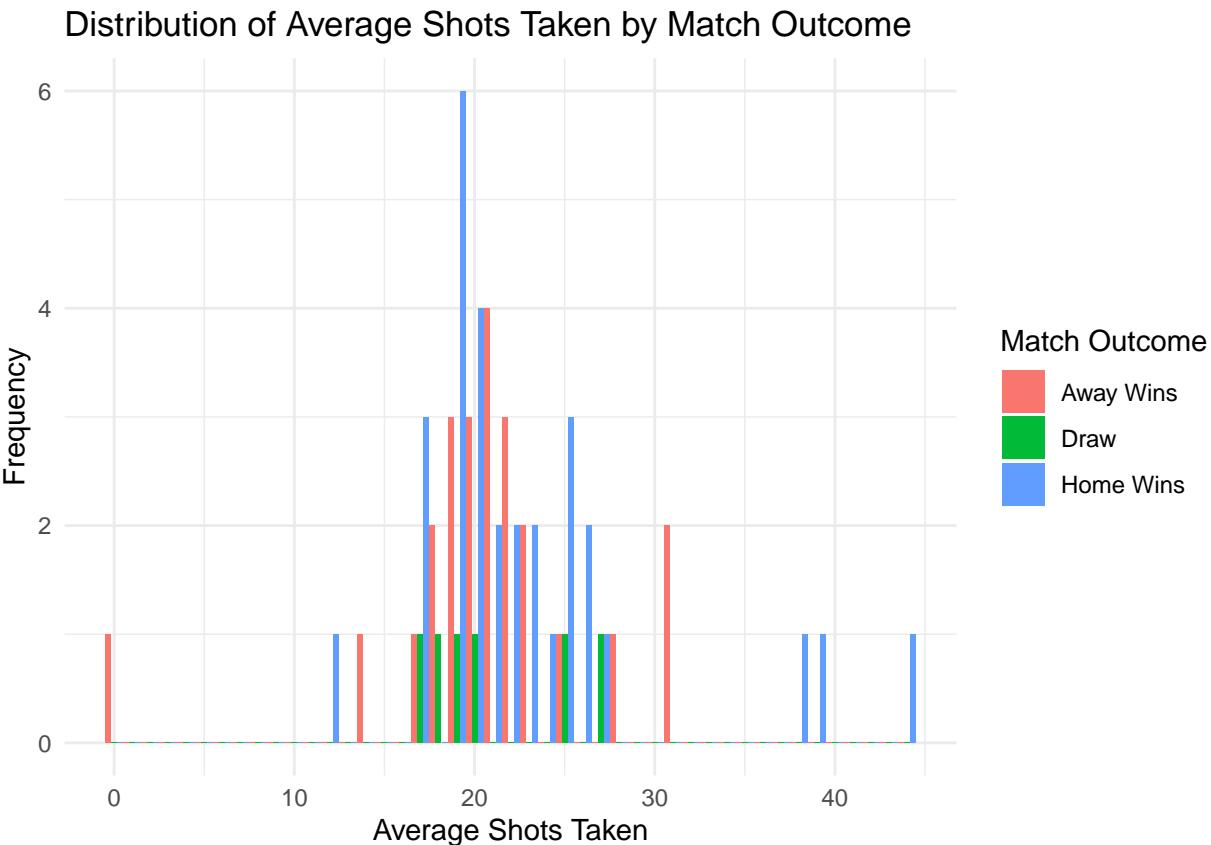
```

```

summarise(AvgShots = mean(HomeTeamShots + AwayTeamShots),
          AvgShotsOnTarget = mean(HomeTeamShotsOnTarget + AwayTeamShotsOnTarget),
          .groups = "drop") %>%
mutate(Result = case_when(
  FullTimeHomeTeamGoals > FullTimeAwayTeamGoals ~ "Home Wins",
  FullTimeHomeTeamGoals < FullTimeAwayTeamGoals ~ "Away Wins",
  TRUE ~ "Draw"
))

# Plotting average shots taken using histograms
ggplot(match_summary, aes(x = AvgShots, fill = Result)) +
  geom_histogram(binwidth = 1, position = "dodge") +
  labs(title = "Distribution of Average Shots Taken by Match Outcome",
       x = "Average Shots Taken",
       y = "Frequency",
       fill = "Match Outcome") +
  theme_minimal()

```

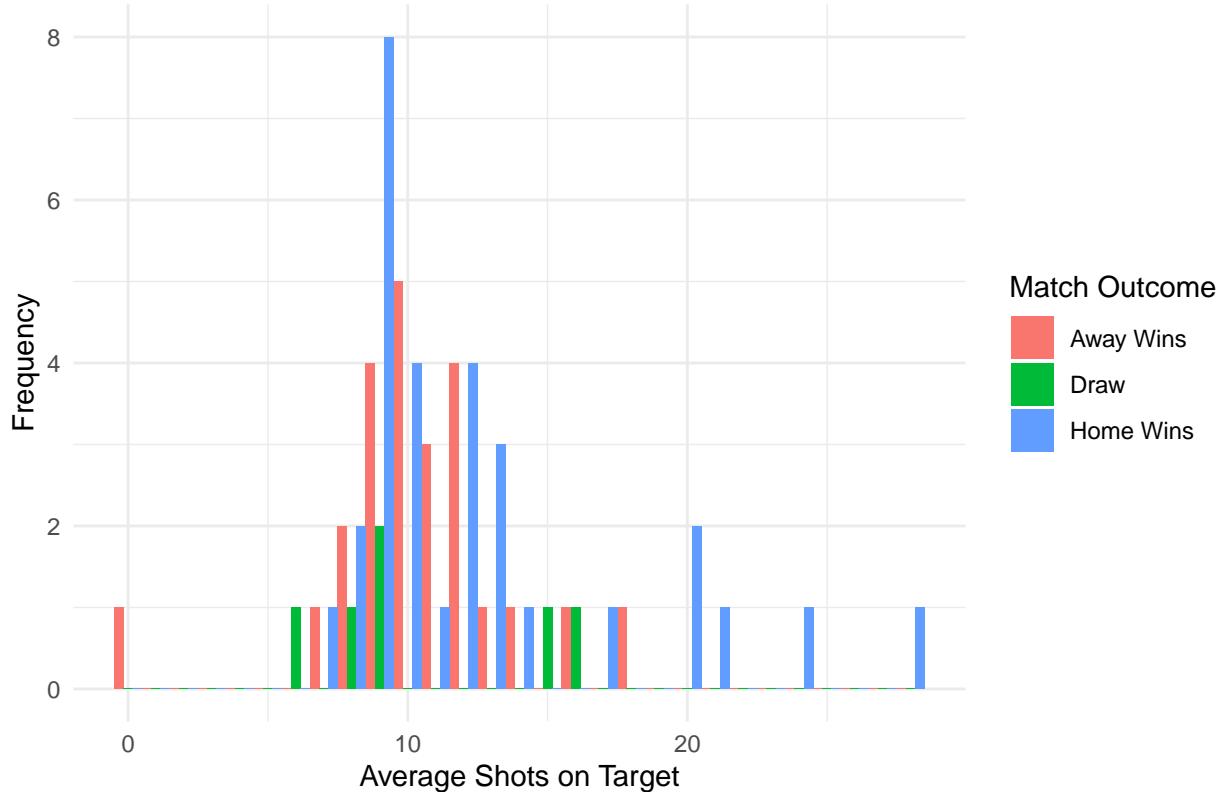


```

# Plotting average shots on target using histograms
ggplot(match_summary, aes(x = AvgShotsOnTarget, fill = Result)) +
  geom_histogram(binwidth = 1, position = "dodge") +
  labs(title = "Distribution of Average Shots on Target by Match Outcome",
       x = "Average Shots on Target",
       y = "Frequency",
       fill = "Match Outcome") +
  theme_minimal()

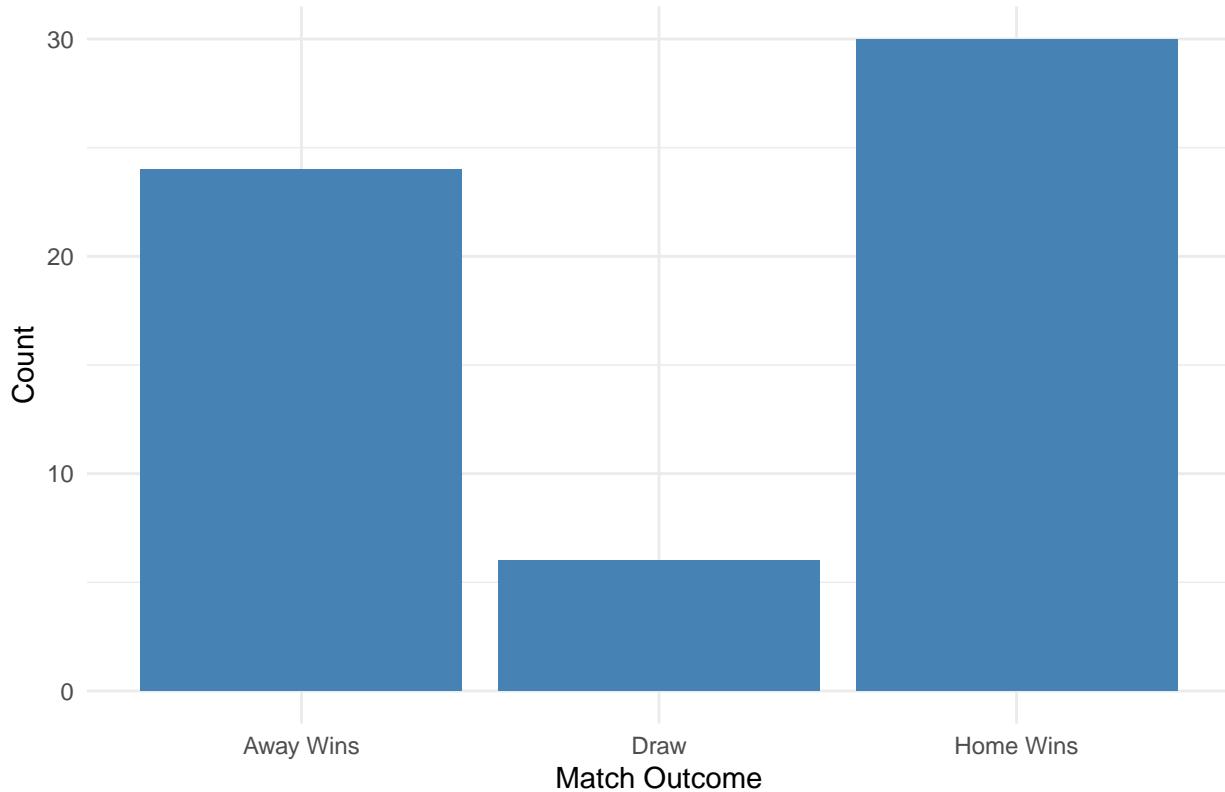
```

Distribution of Average Shots on Target by Match Outcome



```
# Plotting match outcomes
ggplot(match_summary, aes(x = Result)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Distribution of Match Outcomes",
       x = "Match Outcome",
       y = "Count") +
  theme_minimal()
```

Distribution of Match Outcomes



We see that the number of total shots taken are almost the same by both the teams so this might not be the best metric to go by. But we do see a pattern that shows us that the team which has more shots on target wins most number of times so shots on target might be a good metric to go by.

Conclusion

With all the knowledge we have now, we can show that what is the parameter that attributes to winning the most. We can quantify with the findings that we have for example we saw that the number of red cards was a good metric to see which team won and which team lost. The second one was to see how the number of shots and shots on target attributed to winning. With some parameters being more important than the other it was hard to identify which team won and which team lost.

```
# Data cleaning and preprocessing
# Fill missing values with median
data <- data %>%
  mutate(across(everything(), ~ifelse(is.na(.), median(., na.rm = TRUE), .)))

# Create binary outcome variable: 1 if home team wins, 0 otherwise
data <- data %>%
  mutate(Win = ifelse(FullTimeHomeTeamGoals > FullTimeAwayTeamGoals, 1, 0))

# Select relevant predictor variables
predictors <- data %>%
  select(HomeTeamShotsOnTarget, HomeTeamShots, HomeTeamRedCards, AwayTeamShotsOnTarget, AwayTeamShots, ...)

# Plot scatterplot matrix with linear regression line
scatterplot <- predictors %>%
```

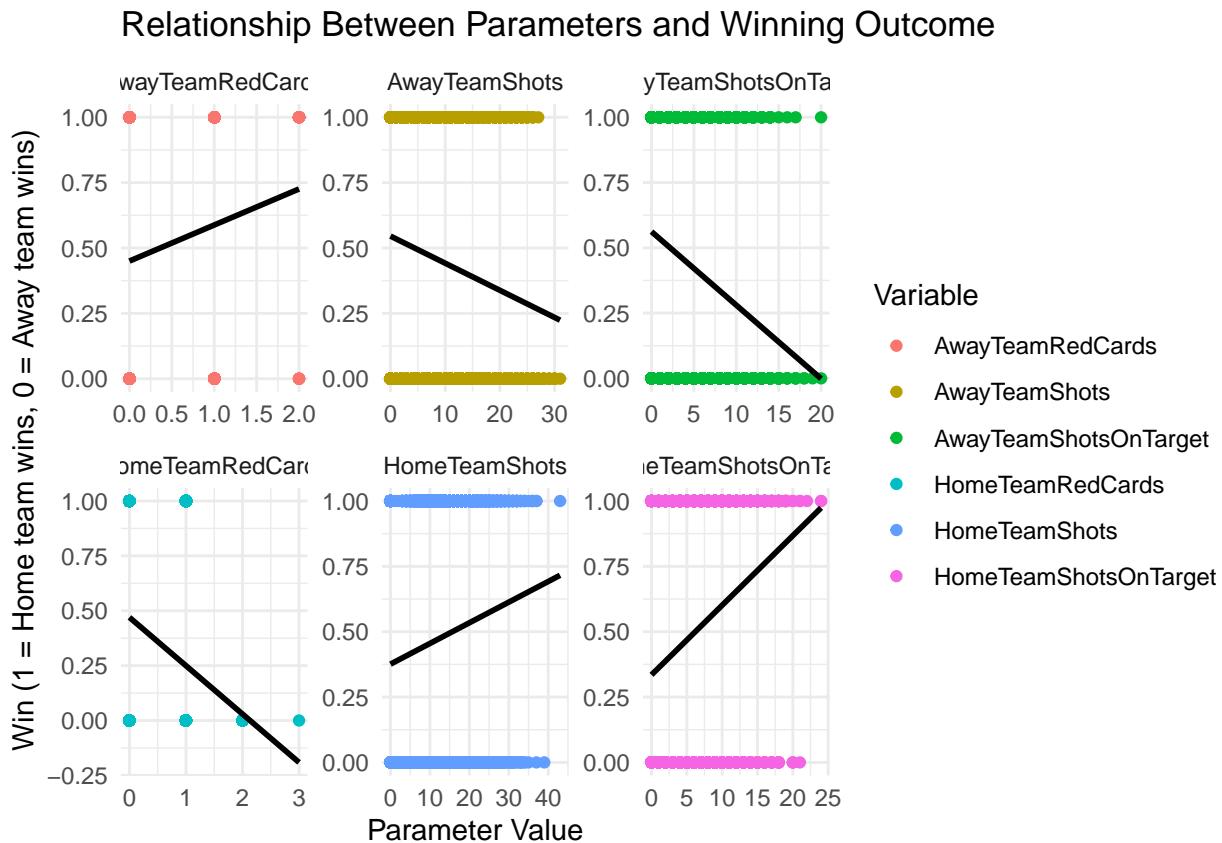
```

gather(key = "Variable", value = "Value", -Win) %>%
ggplot(aes(x = Value, y = Win, color = Variable)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE, color = "black") +
facet_wrap(~Variable, scales = "free") +
labs(title = "Relationship Between Parameters and Winning Outcome",
x = "Parameter Value",
y = "Win (1 = Home team wins, 0 = Away team wins)") +
theme_minimal()

# Display the scatterplot
print(scatterplot)

```

`geom_smooth()` using formula = 'y ~ x'



Final Conclusion

Based on the linear regression model and the scatterplot matrix analysis, we can draw several conclusions regarding the factors that influence the outcome of Premier League matches:

Shots on Target: The number of shots on target, both by the home and away teams, appears to be a significant predictor of match outcomes. A higher number of shots on target tends to correlate positively with the likelihood of winning for both home and away teams.

Total Shots Taken: While the total number of shots taken by each team does not show a clear linear relationship with match outcomes, it seems that having more shots on target is more important than simply having a higher volume of shots overall.

Red Cards: The presence of red cards, whether received by the home or away team, seems to have a discernible impact on match outcomes. Matches where one team receives a red card tend to favor the opposing team, suggesting that red card incidents significantly influence the flow and dynamics of the game.

Home Advantage: While not explicitly analyzed in the scatterplot matrix, the concept of home advantage is well-documented in football. Matches played at the home team's stadium often result in a higher probability of winning for the home team. This factor likely interacts with other variables, such as shots on target and red card incidents, to influence match outcomes.

In conclusion, while various factors contribute to the outcome of Premier League matches, including shots on target, red card incidents, and the home advantage, the number of shots on target emerges as a particularly strong predictor of match outcomes. Teams that can consistently generate quality scoring opportunities by hitting the target are more likely to secure victories, regardless of whether they are playing at home or away. Additionally, the presence of red card incidents introduces an element of unpredictability, often tilting the balance of the game in favor of the opposing team. Therefore, teams should prioritize disciplined play and effective shot execution to improve their chances of winning in the Premier League.

Work Cited

Data Set by IVANRAMOSDATATECH on Kaggle.com <https://www.kaggle.com/datasets/ajaxianazarenka/premier-league>