EPL Analysis

Faizan Husain, Luke Ruan, Akhil Thakur, Neethi Thangiah

16 September 2021

Intoduction

The English Premier league or the EPL is the top level of the English football league system . The premier league is the most watched sports league in the world, broadcasted to 643 million homes. It has one of the most match attendance, second only to Bundesliga. In this markdown, We try to find the factors which influence the most on the teams chance for winning a match.

Dataset

Loading the required library

```
library(tidyverse)
```

Loading the dataset

We used the EPL data set sourced from Kaggle. We directly load the data set from the github repository into the data frame.

```
url <- "https://raw.githubusercontent.com/Neethithevan/EPL-Data-Analysis/main/Dataset/"

csv_names = c("2000-01.csv","2001-02.csv","2002-03.csv","2003-04.csv","2004-05.csv",
              "2005-06.csv","2006-07.csv","2007-08.csv","2008-09.csv","2009-10.csv",
              "2010-11.csv","2011-12.csv","2012-13.csv","2013-14.csv","2014-15.csv",
              "2015-16.csv","2016-17.csv","2017-18.csv","2018-19.csv","2019-20.csv")

url_csv <- str_c(url , csv_names)

data_all <- url_csv %>%
              lapply(read_csv, show_col_types = FALSE) %>% bind_rows
```

```
df <- data_all %>% select("Date","HomeTeam","AwayTeam","FTHG","FTAG","FTR",
                          "HTHG","HTAG","HTR","HST","AST","HF","AF", "HS",
                          "AS", "Referee","HY", "AY", "HR", "AR","Attendance")
head(as_tibble(df))
```

```
## # A tibble: 6 x 21
##   Date    HomeTeam AwayTeam  FTHG  FTAG FTR    HTHG  HTAG HTR     HST   AST    HF
##   <chr>  <chr>    <chr>    <dbl> <dbl> <chr> <dbl> <dbl> <chr> <dbl> <dbl> <dbl>
## 1 19/08~ Charlton Man City     4     0 H         2     0 H        14     4    13
## 2 19/08~ Chelsea  West Ham     4     2 H         1     0 H        10     5    19
## 3 19/08~ Coventry Middles~     1     3 A         1     1 D         3     9    15
## 4 19/08~ Derby    Southam~     2     2 D         1     2 A         4     6    11
## 5 19/08~ Leeds    Everton      2     0 H         2     0 H         8     6    21
## 6 19/08~ Leicest~ Aston V~     0     0 D         0     0 D         4     3    12
## # ... with 9 more variables: AF <dbl>, HS <dbl>, AS <dbl>, Referee <chr>,
## #   HY <dbl>, AY <dbl>, HR <dbl>, AR <dbl>, Attendance <dbl>
```

Exploring the data set

```
sprintf("The number of rows : %d",dim(df)[1])
```

```
## [1] "The number of rows : 7260"
```

```
sprintf("The number of columns : %d", dim(df)[2])
```

```
## [1] "The number of columns : 21"
```

```
team_names <- df %>% distinct(HomeTeam)
sprintf("Total Number of Teams : %d", dim(team_names)[1])
```

```
## [1] "Total Number of Teams : 44"
```

Analysis

Analysis 1 : Goals Average/ Game through the 20 seasons

```
season_vec <- c("2000-01","2001-02","2002-03","2003-04","2004-05","2005-06",
                "2006-07","2007-08","2008-09","2009-10","2010-11","2011-12",
                "2012-13","2013-14","2014-15","2015-16","2016-17","2017-18",
                "2018-19","2019-20")
season_col <- c()

for(i in 1:20){
  if(i<19){season_col <-  c(season_col,rep(season_vec[i],380))}
  else if( i == 19){season_col <-  c(season_col,rep(season_vec[i],160))}
  else if(i==20){season_col <-  c(season_col,rep(season_vec[i],260))}
}

df$Season <- season_col # Adding it to the date frame df
```

We now calculate the average number of goal scored in the league through the seasons.

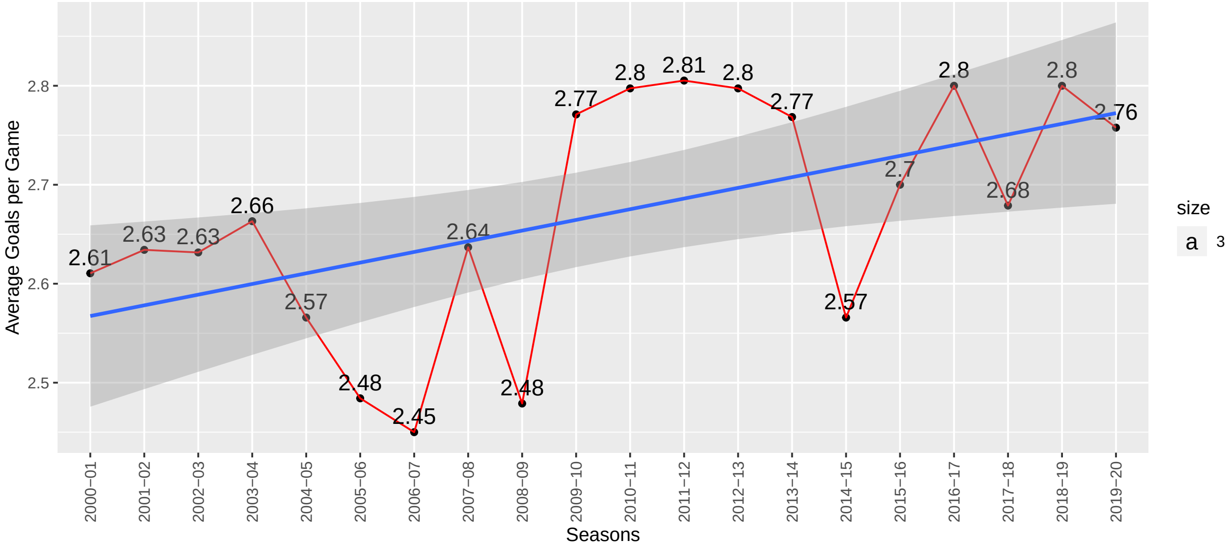
```
# Finding the goal average per match/game and summarizing the result in season wise
df_goal_avg_season <- df %>%mutate(TG = FTHG + FTAG) %>%
  group_by(Season) %>%
  summarise(goal_avg = sum(TG)/n())
head(df_goal_avg_season)
```

```
## # A tibble: 6 x 2
##   Season  goal_avg
##   <chr>      <dbl>
## 1 2000-01     2.61
## 2 2001-02     2.63
## 3 2002-03     2.63
## 4 2003-04     2.66
## 5 2004-05     2.57
## 6 2005-06     2.48
```

```
# Plotting the results in graph
ggplot(df_goal_avg_season, aes(Season ,goal_avg ,group = 2)) +
  geom_point()+
  geom_line(col = "red" ,bg = "blue")+
  geom_text(aes(label=signif(goal_avg ,3), size= 3 , vjust = -0.5)) +
  geom_smooth(method = "lm")+
  ylab("Average Goals per Game")+
  xlab("Seasons")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
## Warning: Ignoring unknown parameters: fill
```

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



Analysis 1 Inference

We can see an increasing trend in the average number of goals scored from season to season.

Analysis 2 : Home Ground Advantage

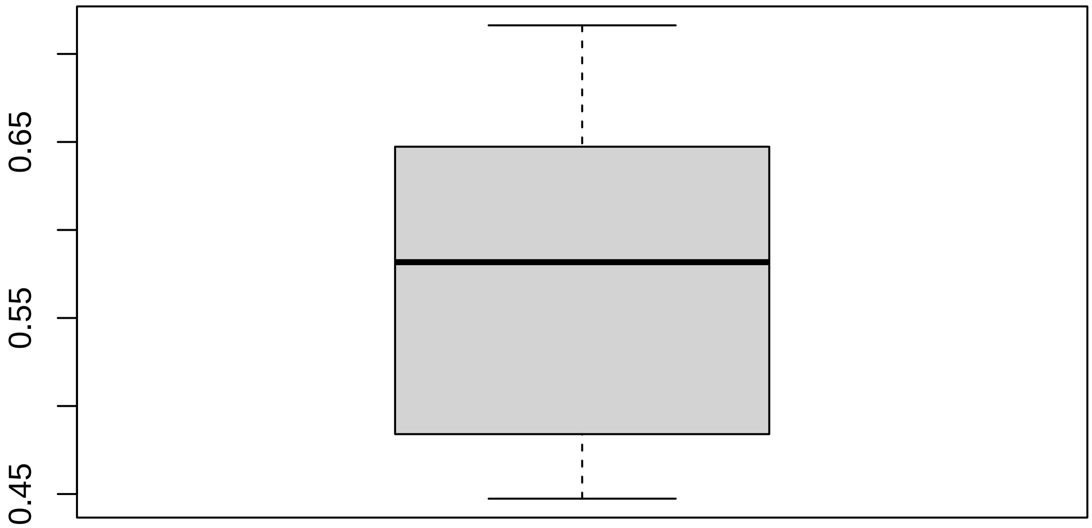
```
df %>% group_by(FTR) %>% summarize(number_of_wins = n()) %>% mutate(percent = number_of_wins/sum(number_
```

```
## # A tibble: 3 x 3
##   FTR   number_of_wins percent
##   <chr>          <int>   <dbl>
## 1 A               2049   0.282
## 2 D               1851   0.255
## 3 H               3360   0.463
```

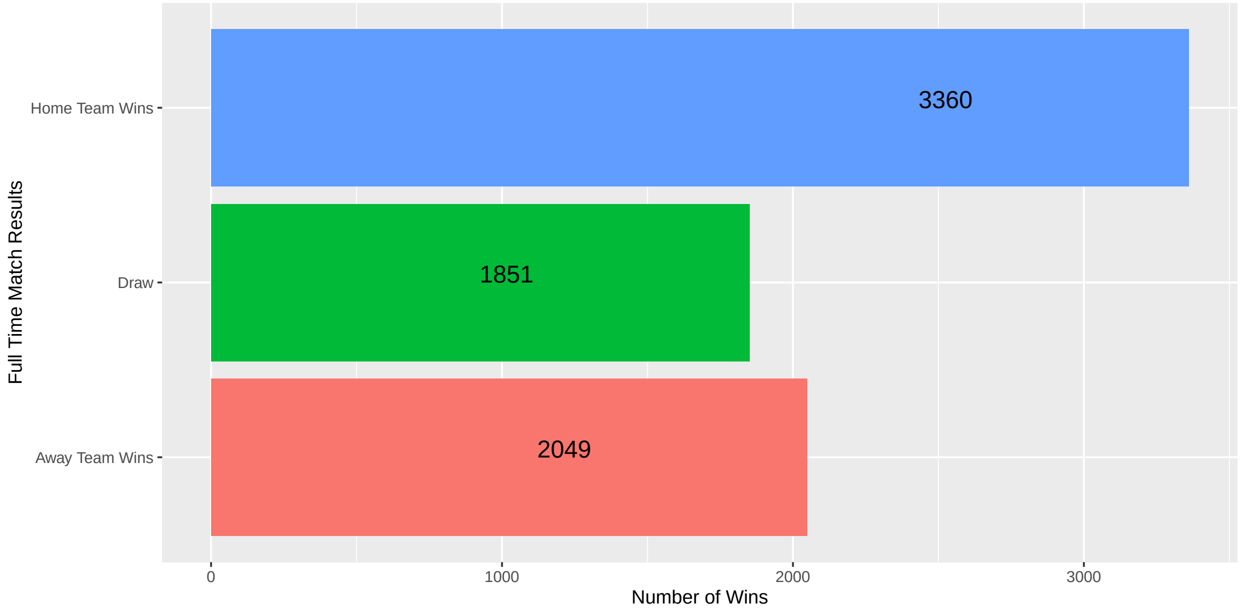
```
df_home_ground<-df %>% group_by(HomeTeam,FTR) %>% summarise(wins= n()) %>%
  ungroup() %>% group_by(HomeTeam) %>%  mutate(total_wins = sum(wins)) %>%
  ungroup() %>% mutate(win_percentage= wins/total_wins) %>% filter(FTR == "H") %>%
  arrange(desc(win_percentage)) %>% head(11)
```

```
## ‘summarise()‘ has grouped output by ’HomeTeam’. You can override using the ‘.groups‘ argument.
```

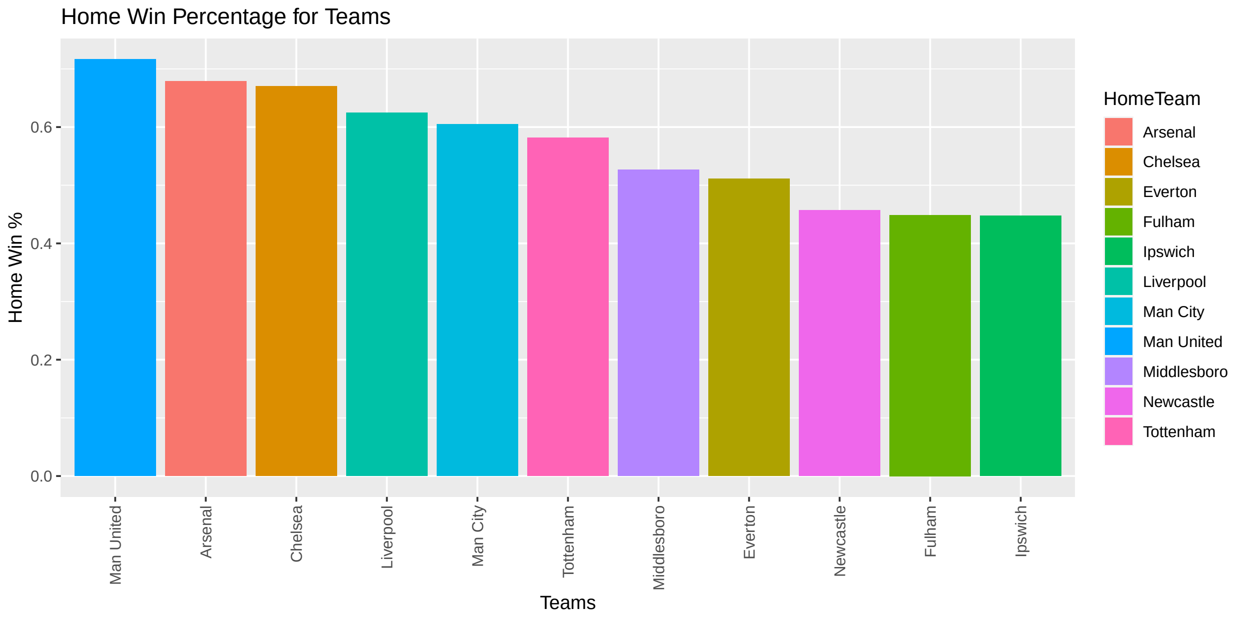
```
boxplot(df_home_ground$win_percentage)
```



```
ggplot(df %>% group_by(FTR) %>% summarize( number_of_wins = n()) ,
       aes(x = c("Away Team Wins", "Draw", "Home Team Wins") , y = number_of_wins,fill = FTR))+
  geom_bar(stat ="identity") + coord_flip() +theme_grey() +
  geom_text(aes(label=number_of_wins), vjust=0 , hjust = 5 ,size = 5)+
  ylab("Number of Wins")+
  xlab("Full Time Match Results")+
  theme(legend.position = "none")
```



```
## Analyzing Home Wins by teams

ggplot(df_home_ground , aes(x = reorder(HomeTeam ,-win_percentage), y = win_percentage ,fill = HomeTeam
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  xlab("Teams")+
  ylab("Home Win %")+
  ggtitle( "Home Win Percentage for Teams ")
```

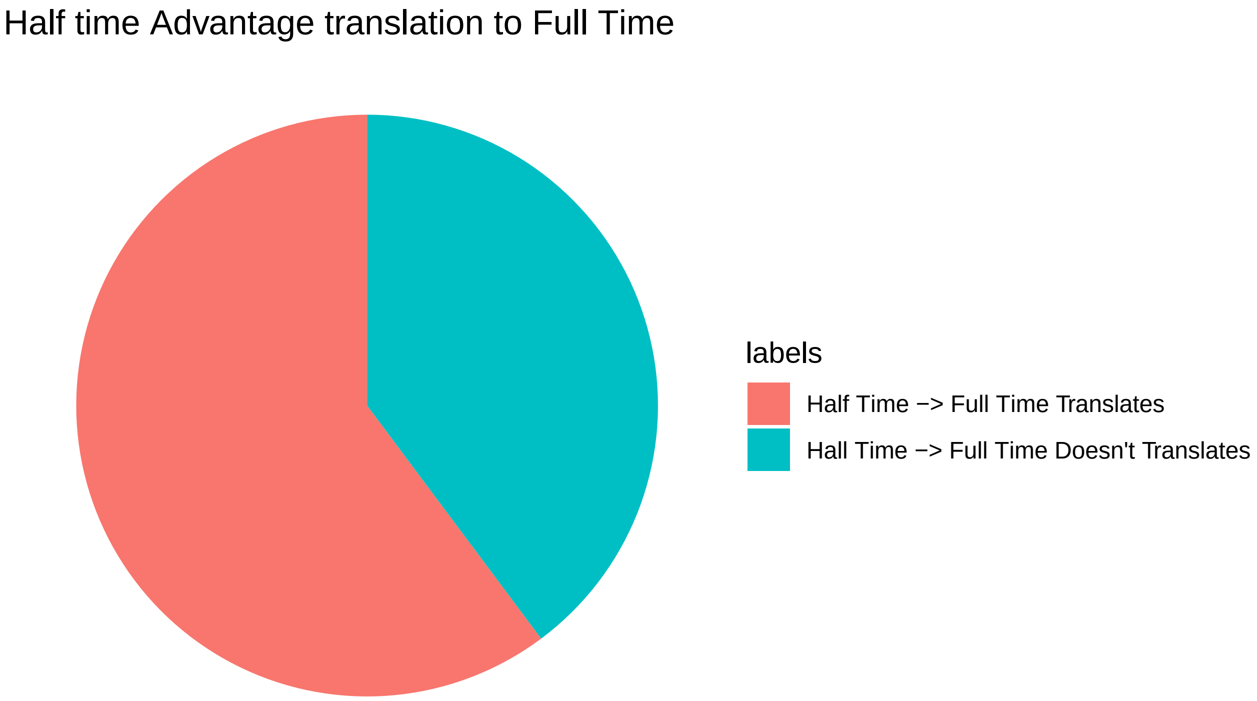


Analysis 2 Inference

- Number of wins on home field significantly outnumbers both losses and draws.
- Median winning percentage on home field is >55%.
  - Therefore there seems to be a correlation between playing on home field and a higher win percentage

Analysis 3 : Do Halftime Leads Translate to Full-time Wins ?

```
## Hypothesis : Result at the end of first half will might well be the final result

df_htr <- cbind(df  %>%filter(HTR == FTR) %>% summarise(HTR_translates_FTR =n()),
                df %>% filter(HTR != FTR) %>% summarize(HTR_doesnt_translate_FTR = n()))
df_htr_pie = data.frame(x <- c(df_htr$HTR_translates_FTR, df_htr$HTR_doesnt_translate_FTR),
                labels <- c("Half Time -> Full Time Translates", "Hall Time -> Full Time Doesn't Transla
ggplot(df_htr_pie, aes(x="", y=x, fill=labels)) +geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +theme_void() + ggtitle("Half time Advantage translation to Full Time")
```



```
df_htr <- df_htr %>%
  mutate(HTR_translates_FTR_per = HTR_translates_FTR/(HTR_translates_FTR+HTR_doesnt_translate_FTR))
df_htr
```

```
##   HTR_translates_FTR HTR_doesnt_translate_FTR HTR_translates_FTR_per
## 1               4371                     2889              0.6020661
```

Analysis 3 Inference

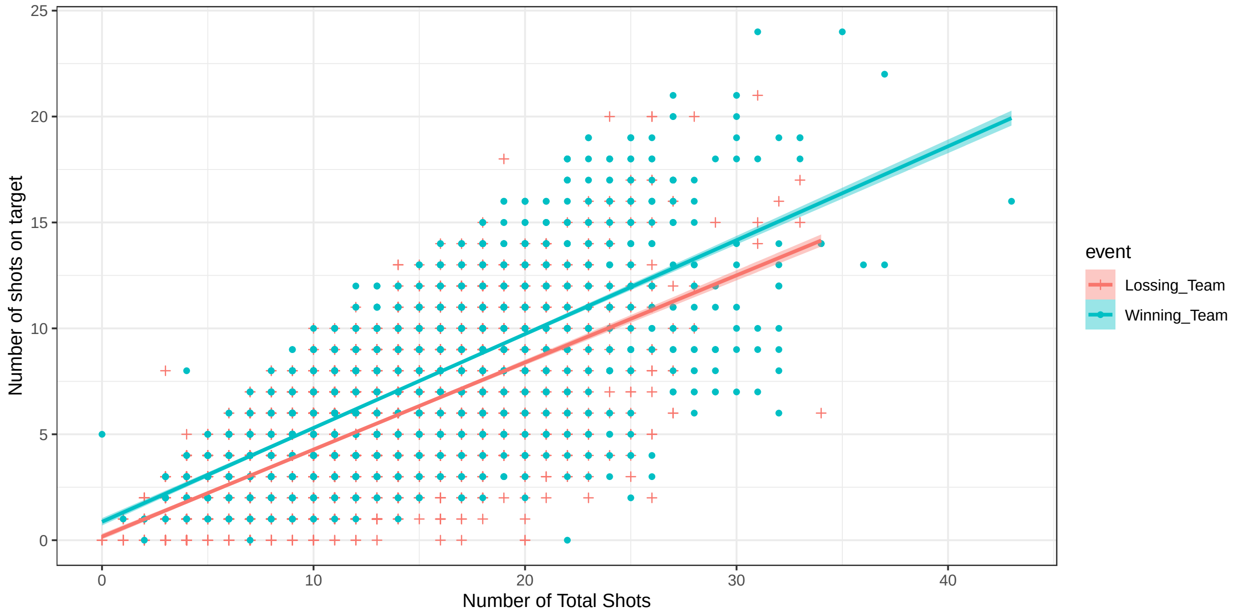
- Most of the time, winning team at halftime will win the game.
- ~40% of the time, the losing team at halftime will come back to win

Analysis 4 : Shots Taken vs. Shots On Target

```
df_st_h <- df %>% filter(FTR =='H' ) %>% select(FTHG,HST, HS)
colnames(df_st_h)<- c('GoalsScored','ST', 'S')

df_st_a <- df %>% filter(FTR =='A' ) %>% select(FTAG,AST, AS)
colnames(df_st_a)<- c('GoalsScored','ST', 'S')

df_st_w <- rbind(df_st_a ,df_st_h)
df_st_w$event <- "Winning_Team"
```

```
#For Lossing Team

df_st_h <- df %>% filter(FTR =='H' ) %>% select(FTAG,AST, AS)
colnames(df_st_h)<- c('GoalsScored','ST', 'S')

df_st_a <- df %>% filter(FTR =='A' ) %>% select(FTHG,HST, HS)
colnames(df_st_a)<- c('GoalsScored','ST', 'S')

df_st_l <- rbind(df_st_a ,df_st_h)

df_st_l$event <- "Lossing_Team"

df_st <- rbind(df_st_l , df_st_w)
```

```
#Building a linear model
ggplot(df_st , aes(x= S ,y = ST,colour = event , fill = event)) +
  geom_point(aes(shape=event)) +
  scale_shape_manual(values=c(3,16))+
  geom_smooth(method = "lm")+
  theme_bw() +
  xlab("Number of Total Shots") +
  ylab("Number of shots on target")
```

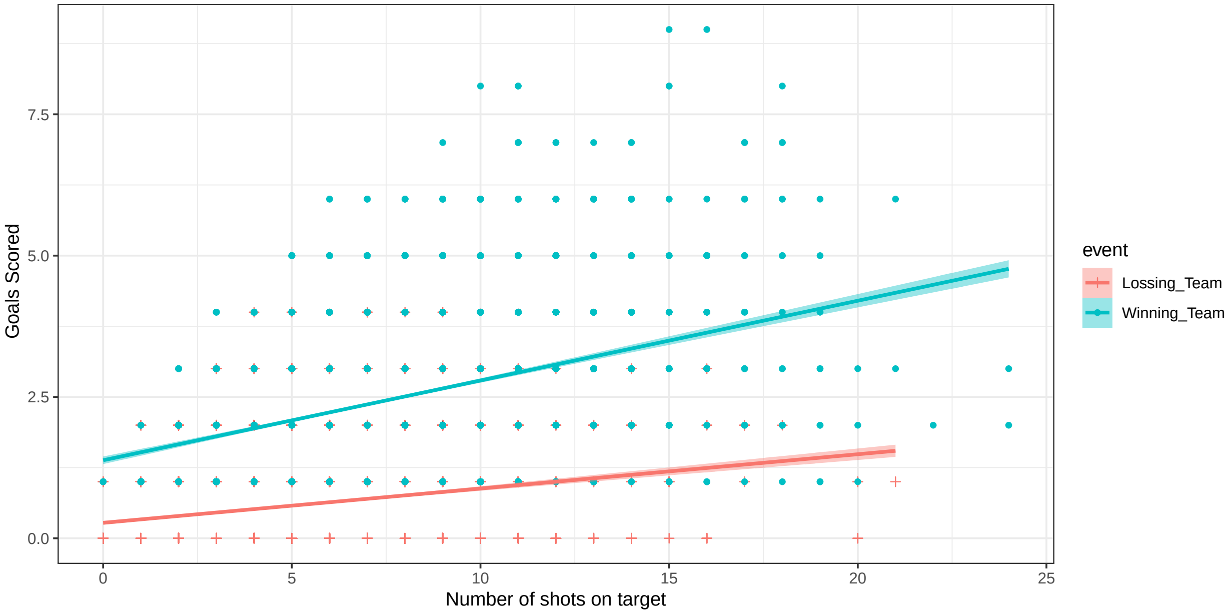


Analysis 4 Inference

- Positive correlation between Shots Taken and Shots On Target.
  - Take more Shots to score more Shots on Target.

Analysis 5 : Shots On Target vs. Goals Scored

```
#Building a linear model
ggplot(df_st , aes(x= ST ,y = GoalsScored,colour = event , fill = event)) +
  geom_point(aes(shape=event)) +
  scale_shape_manual(values=c(3,16))+
  geom_smooth(method = "lm")+
  theme_bw() +
  xlab("Number of shots on target") +
  ylab(" Goals Scored")
```



Analysis 5 Inference

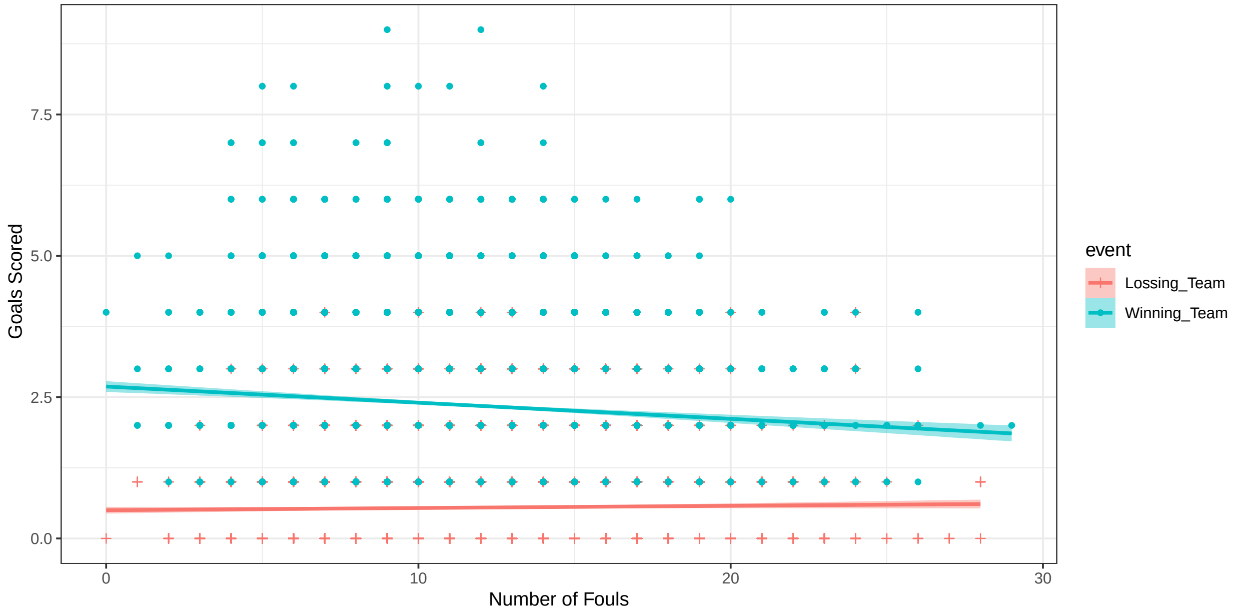
- Positive correlation between both:
- Shots Taken and Shots On Target
- Shots On Target and Goals Scored
    - Therefore, take more shots!

Analysis 6 : Does committing more fouls correlate with scoring goals?

```
df_fouls_h <- df %>% filter(FTR =='H' ) %>% select(FTHG,HF,Attendance)
colnames(df_fouls_h)<- c('GoalsScored','Fouls', 'Attendance')

df_fouls_a <- df %>% filter(FTR =='A' ) %>% select(FTAG,AF, Attendance)
colnames(df_fouls_a)<- c('GoalsScored','Fouls', 'Attendance')

df_fouls_w <- rbind(df_fouls_a ,df_fouls_h)
df_fouls_w$event <- "Winning_Team"
```

```
#For Lossing Team

df_fouls_h <- df %>% filter(FTR =='H' ) %>% select(FTAG,AF, Attendance)
colnames(df_fouls_h)<- c('GoalsScored','Fouls', 'Attendance')

df_fouls_a <- df %>% filter(FTR =='A' ) %>% select(FTHG,HF, Attendance)
colnames(df_fouls_a)<- c('GoalsScored','Fouls', 'Attendance')

df_fouls_l <- rbind(df_fouls_a ,df_fouls_h)

df_fouls_l$event <- "Lossing_Team"

df_fouls <- rbind(df_fouls_l , df_fouls_w)
```

```
ggplot(df_fouls , aes(x= Fouls ,y = GoalsScored,colour = event , fill = event)) +
  geom_point(aes(shape=event)) +
  scale_shape_manual(values=c(3,16))+
  geom_smooth(method = "lm")+
  theme_bw() +
  xlab("Number of Fouls") +
  ylab(" Goals Scored")
```



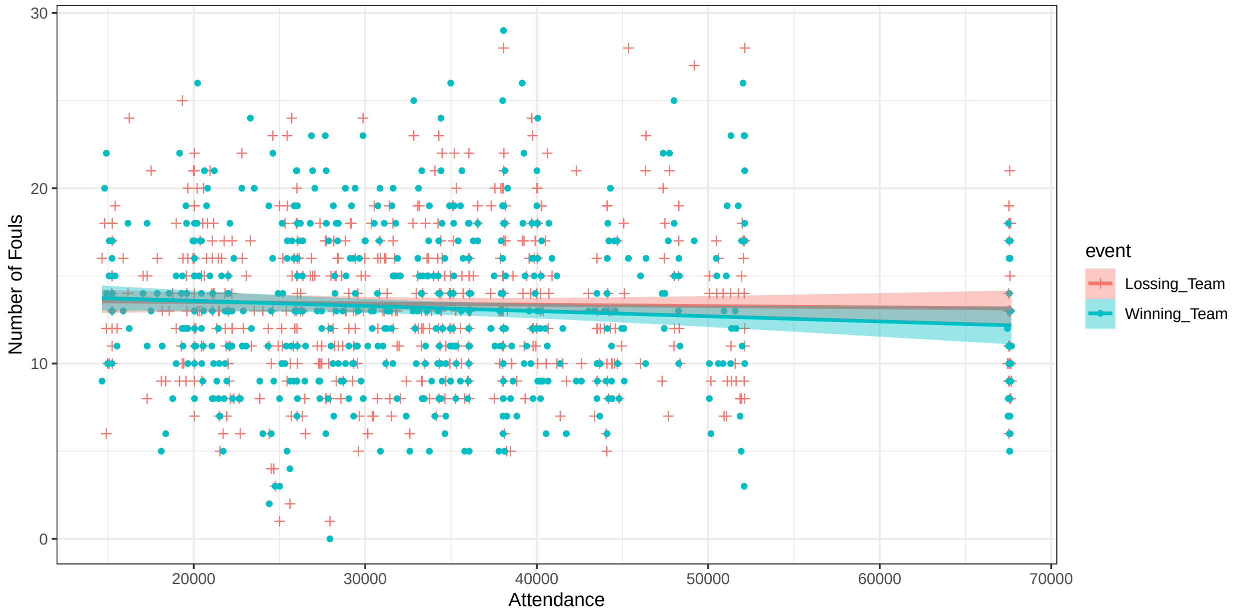
Analysis 6 Inference

- Negative correlation between fouls committed and goals scored for the winning team
  - Therefore, commit less fouls

Analysis 7 : Audience Vs Fouls committed.

Let's see how the audience affect the gameplay. Specifically, lets see if the attendance affects the number of fouls committed by the players.

```r
ggplot(df_fouls , aes(x= Attendance ,y = Fouls,colour = event , fill = event)) +
  geom_point(aes(shape=event)) +
  scale_shape_manual(values=c(3,16))+
  geom_smooth(method = "lm")+
  theme_bw() +
  xlab("Attendance") +
  ylab("Number of Fouls")
```



Analysis 7 Inference

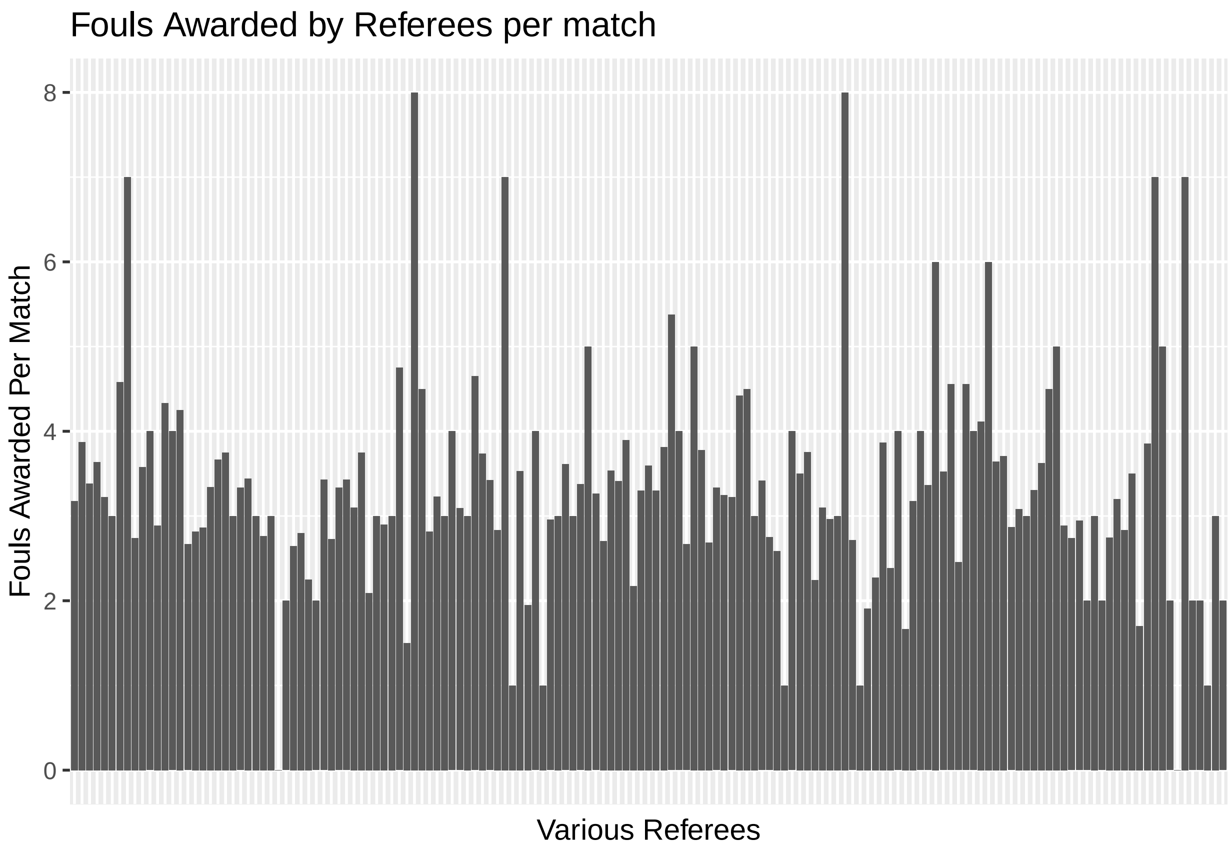
- We see that there is no significant affect on the happening of fouls due to the attendance.

BIAS

- We only looked at EPL. The factors affecting the match might not be significant in other leagues.
- We looked at the entire past 20 of data for out inferences. There might be significant changes in the gameplay in the recent years compared to say last decade.
- There might be some other factors affecting the gameplay of the match, such as the fouls awarded by the referee. So, let see it that's actually the case.

Are some referees more inclined to give out more fouls?

```r
df_ref <- df %>% group_by(Referee) %>%
  summarize(count = n(), fouls=sum(HR+AR+AY+HY)) %>%
  ungroup()

df_ref %>% group_by(Referee) %>%
  ggplot(aes(x = Referee, y = fouls/count)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  xlab("Various Referees") +
  ylab("Fouls Awarded Per Match") +
  ggtitle( "Fouls Awarded by Referees per match") +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



Inference

- We can see that some of the referees award more fouls on average
- So, there is a chance that there might be some bias introduced by the referee.

Conclusion

Formula For Victory.

Based on our above analysis, we can conclude that following this formula will lead to victory.

1. Play on your home field!

2. Get an early lead
3. Take more shots
4. Commit fewer (or smarter) fouls