# Homework 3

*Narek Sahakyan*

*27 October, 2019*

TOTAL: 50p

Reminder: In football draw is a half win. Number of teams per season might change even within given league

Noll-Skully number

*Your goal is to calculate Noll-Skully number for football* Chose any 4 leagues from f_data_sm on your own * Calculate Noll_Skully Number as a whole for all the seasons for the given league. (2p)

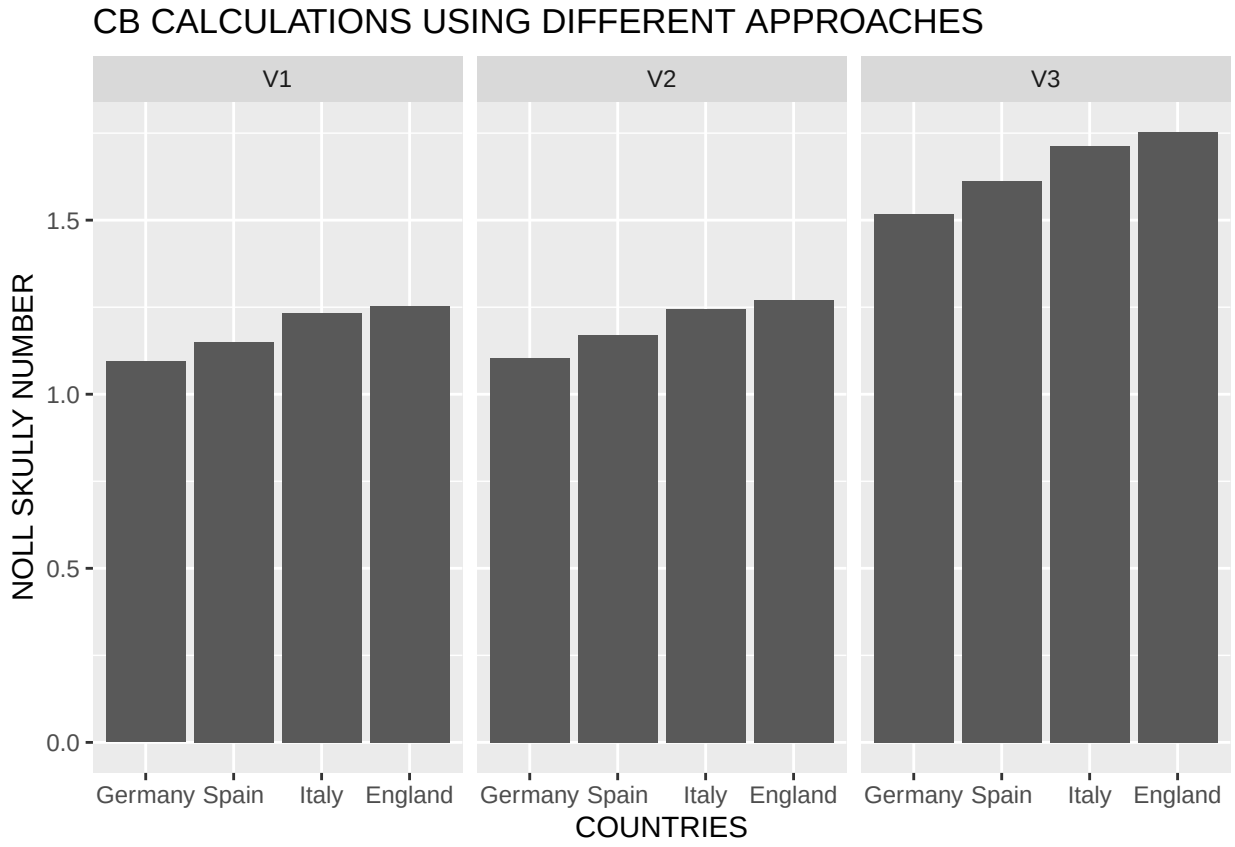
```
top_4_data <- f_data_sm %>%
  filter(COUNTRY %in% countries)

final_tables <- function(data,country){
  result <- c()
  for(season in unique(data$SEASON)){
    season_table <- final_table(data, country, season)
    league <- data[data$SEASON == season &
                     data$COUNTRY == country,]$LEAGUE[1]
    season_table$SEASON = season
    season_table$COUNTRY = country
    season_table$LEAGUE = league
    result <- rbind(result, season_table)
  }
  cols <- colnames(result)
  len <- length(cols)
  result <- result[c("SEASON","COUNTRY","LEAGUE",cols[1 : (len-3)])]
  return(as.data.frame(result))
}

countries_standings <- function(data){
  result <- c()
  for(country in unique(data$COUNTRY)){
    country_standings <- final_tables(data,country)
    result <- rbind(result, country_standings)
  }
  return(result)
}

top_4_standings <- countries_standings(top_4_data)
top_4_standings$TW = round(top_4_standings$W + (top_4_standings$D)/2)
top_4_standings$WR <- top_4_standings$TW / top_4_standings$M
#Win Ratio = number of wins / number of games
#Tunned wins = number of wins + (number of draws)/2
#As already mentioned in the description the number of games played
#each season in football can change even in the same league,
#so in order to have a general metric for summarizing let's take
#the mean of the number of games played in each season as an approximation
```

```r
#We will use the exact number of games later,
#when analyzing the CB using season based approaches

#Version 1 teams_count
#Number of teams = number of games / 2 + 1
# teams_count <- top_4_standings %>%
#   group_by(COUNTRY) %>%
#   summarise(MEAN.TC = round(mean( M/2 + 1 )))
# id_s <- 0.5 / sqrt(teams_count)

mean_t <- function(M) {
  return(round( mean( M/2 + 1 ) ))
}

top_4_nsc_v1 <- top_4_standings %>%
  group_by(COUNTRY) %>%
  summarise(NSC = sd(WR) / ( 0.5 / sqrt(mean_t(M)))) %>%
  mutate(V = "V1") %>%
  arrange(desc(NSC))
#Now let's use count three draws as one win
top_4_standings$TW_v2 <- round(top_4_standings$W + (top_4_standings$D)/3)
top_4_standings$WR_v2 <- top_4_standings$TW_v2 / top_4_standings$M
top_4_nsc_v2 <- top_4_standings %>%
  group_by(COUNTRY) %>%
  summarise(NSC = sd(WR_v2) / ( 0.5 / sqrt(mean_t(M)))) %>%
  mutate(V = "V2") %>%
  arrange(desc(NSC))

# As we can see the competitive balance increased for all the leagues
# when using this type of calculations.
# The increase in the competetive balance
# means that the role of luck gets less decisive.
# However the changes are very slight and
# not significant, so I am not sure,
# but I believe that having these facts we can conclude
# that the draws in general do slightly affect the all time CB for these leagues, as
# when we give them less importance(v1: 2 draws = 1 win, v2: 3 draws = 1 win)
# the CB is increasing which can be translated into
# the decrease of the importance of the luck. I find this connection logical
# as if the CB is getting higher when we discard more draws,
# as we count 3 of them as a win. However when the CB is close to
# being balanced we expect more draws in general.
# Now let's imagine two simple examples.
# Suppose team X won 1 game and tied 10 times.
# If we want to transform the draws into wins using half principle
# We would say that X won 6 of its games, using 1/3 principle
# we would say that the team won 4 of its games.
# According to our approach, in general the discard of that two "won" games leads to
# increase in CB(decrease in luck importance).
# Using 1/3 principle the teams with most wins will get the
# highest portion of wins, and the teams with more draws
# will get lower rates in comparison to the rates
# they would get using half principle(6 vs 4) and
```

```r
# this will lead to higher variance in skill(CB). So the teams
# with draws the lower will become this variance in skill,
# which will increase the luck's importance as in
# general there would be more equal teams and more expected draws.

# I hope I have written something logical :)



#Version 2 matches count
# Let's use the same approach as for version one,
# but considering the number of games played
# rather than the number of the teams.
mean_m <- function(M){
  return(round(mean(M)))
}

top_4_nsc_v3 <- top_4_standings %>%
  group_by(COUNTRY) %>%
  summarise(NSC = sd(WR_v2) / ( 0.5 / sqrt(mean_m(M)))) %>%
  mutate(V = "V3") %>%
  arrange(desc(NSC))

top_4_nsc <- rbind(top_4_nsc_v1,
                   top_4_nsc_v2,
                   top_4_nsc_v3)
# As we can see all the approaches identified the same ranking
# of the leagues in terms of CB over all time, so let's check
# the differences in CB for the leagues for all of the approaches
```

- Plot these numbers together and comment (6p)

```r
top_4_nsc %>%
  ggplot(aes(x = reorder(COUNTRY,NSC))) +
  geom_bar(aes(y = NSC),stat = "identity") +
  facet_grid(.~V) +
  labs(x = "COUNTRIES", y = "NOLL SKULLY NUMBER") +
  ggtitle("CB CALCULATIONS USING DIFFERENT APPROACHES")
```




```
# As we can see the differences in the leagues in terms of CB balances remain
# almost the same for V1 and V2, but become more sparse for V3.
# However this is mainly because of considering
# the number of matches instead of number of teams.
# I believe that using V3 is more relevant for football.
# As we know in general the
# Number of Games in a competition = (Number of teams - 1 ) * 2,
# So addition of one team will lead to two more games, two teams = 4 more games and so on.
# In general, more the games, more the chances of change in CB.
# Supposing that we calculate ID_S by the formula
# 0.5 / sqrt(FACTOR), where FACTOR
# is either 1)the number of games or 2)the number of teams.
# Unit increase in 1) leads to double of that increase in 2),
# the higher the denominator the lower the ID_S
# which in case will lead to higher CB as CB = SD(WPCT)/ID_S.
# As number of games are derived
# from the number of teams(opposite is also true in general) and in general
# they are more important in CB as CB is more sensitive to it and becomes
# higher when regarding them as factor,
# In my opinion it is better to use number of games as a FACTOR for ID_S

top_4_nsc %>%
  group_by(V) %>%
  summarise(SD = sd(NSC)) %>%
  arrange(desc(SD))
```

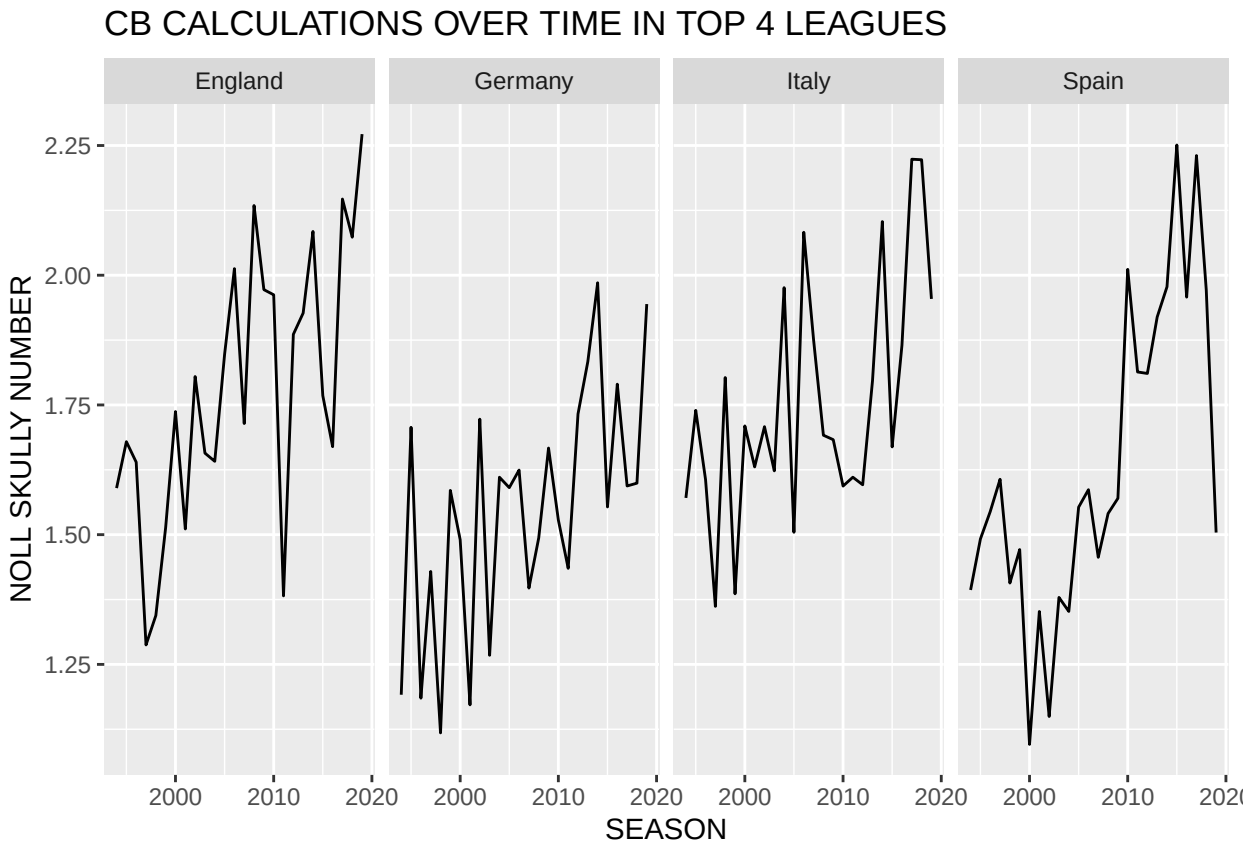
```
## # A tibble: 3 x 2
##   V         SD
##   <chr>  <dbl>
## 1 V3    0.106
## 2 V2    0.0757
## 3 V1    0.0740
```

- Now do the same by season (6p)

```
top_4_nsc_v4 <- top_4_standings %>%
  group_by(SEASON, COUNTRY) %>%
  mutate(WR = (W + D/3) / M) %>%
  summarise(NSN = sd(WR) / (0.5 / sqrt(unique(M)))) %>%
  arrange(desc(NSN))
```

- Plot and comment (8P)

```
highest_cbs <- top_4_nsc_v4 %>%
  group_by(SEASON) %>%
  filter(NSN == max(NSN)) %>%
  arrange(desc(NSN))


narrow_standings <- function(cbs = highest_cbs,standings = top_4_standings){
  result <- c()
  for( index in 1:nrow(cbs) ) {
    competition <- cbs[index,]
    standings <- top_4_standings %>%
      filter(SEASON == competition$SEASON,
             COUNTRY == competition$COUNTRY)
    result <- rbind(result, standings)
  }
  return(result)
}

competitive_standings <- narrow_standings()
# As we take a look at this data frame we can see
# that in these years when the CB was the highest
# the standings were tense, and the difference in points
# for the clubs is minimal for each country. Even in Italy
# in season 2017/2018 Juve was had only 4 more points than
# Napoli. In 1999 the most competitive season was in Germany.
# It is interesting that the difference in points was not that
# tense, this is mainly because Germany had the lowest CB over
# time when not regarding the seasons, meaning that even normal
# differences in points for Germany seem competitive as most of
# the time CB is low.

top_4_nsc_v4 %>%
  ggplot(aes(x = SEASON, y = NSN)) +
  geom_line() +
  facet_grid(.~COUNTRY) +
  labs(x = "SEASON", y = "NOLL SKULLY NUMBER") +
  ggtitle("CB CALCULATIONS OVER TIME IN TOP 4 LEAGUES")
```




```
# As we can see currently and over time England seems to have
# the highest competitive balance. We can see that the highest
# value of CB was in 2019, when Man City won the championship
# by having only one more point than Liverpool. England and Germany
# were only countries in which the season 2018/2019 was more competitive
# than the previous season, whereas In Spain the competitive balance
# rate dropped significantly for season 2018/2019, in Italy the pattern
# was also the same, but it was not that significant and visible as the season
# 2017/2018 was the most competitive.
# In general we can see that the ranking of the leagues in terms of being
# competitive is expressed through the visualization as England is the most competitive
# and Germany has the lowest CB over time in comparison with the others. All of the leagues
# had some peak values(both low and high) and in that period, we can see that the standings
# were accordingly tense or variated.
```

Your goal is to find 2 leading and 2 lagging indicators for those leagues. Show correlation (on plot and calculating correlation coefficient) between these indicators and Noll-Skully number (or any other competitive balance metric on your choice). 20p

Explain why do you think these variables are leading or lagging. 10p

```
# As I have already analyzed Noll-Skully number as
# a CB factor, now let's do calculations for HHI,dHHI
# and C5(top 5 teams)
# As we will play with revenue data later
# let's adopt the standings to them
```

```r
top_4_standings <- top_4_standings %>%
  filter(SEASON >= 1997)

top_4_HHI <- top_4_standings %>%
  group_by(SEASON, COUNTRY) %>%
  mutate(PERC = POINTS / sum(POINTS)) %>%
  summarise(HHI = sum(PERC^2))
top_4_dHHI <- top_4_standings %>%
  group_by(SEASON, COUNTRY) %>%
  mutate(PERC = POINTS / sum(POINTS)) %>%
  summarise(HHI = sum(PERC^2) - 1 / n())

top_5_teams <- top_4_standings %>%
  group_by(SEASON, COUNTRY) %>%
  filter(POSITION <= 5) %>%
  summarise(TOP.P = sum(POINTS))
all_teams <- top_4_standings %>%
  group_by(SEASON, COUNTRY) %>%
  summarise(TOP.P = sum(POINTS))

C5_top_4 <- data.frame(SEASON = top_5_teams$SEASON,
                       COUNTRY = top_5_teams$COUNTRY,
                       C5 = top_5_teams$TOP.P / all_teams$TOP.P)
# Revenue of the teams in the leagues affects the budgets of the teams.
# The budget of the teams define they capability to pay "high" salaries to
# their players. The players earn "high" salary for playing good. So in general
# if the more players get high salary then more players will play good so more
# teams will play good and the competitive balance will probably increase

# First let's look at the deviation in revenue per country
revenue_gap_c <- revenue %>%
  group_by(COUNTRY) %>%
  summarise(TOTAL.REV = sum(REVENUE), SD = sd(REVENUE)) %>%
  arrange(desc(TOTAL.REV))
# English Premier League had the highest revenue over time
# and the highest variation among the league's revenues over
# years.

# The same in years
revenue_gap_s <- revenue %>%
  group_by(SEASON) %>%
  summarise(TOTAL.REV = sum(REVENUE), SD = sd(REVENUE)) %>%
  arrange(desc(TOTAL.REV))
# Absolutely logical order, as nowadays football is
# becoming more and more like an economic field for
# making money, so the revenue and deviation of it in
# top leagues is increasing year by year.

revenue_gap_cs <- revenue %>%
  group_by(COUNTRY,SEASON) %>%
  summarise(TOTAL.REV = sum(REVENUE)) %>%
  arrange(desc(TOTAL.REV))
```

```r
# England! England! England! everywhere
# The main reason behind it is that English
# teams(not only Premier League) earn huge ammounts of
# money from TV contracts compared to other leagues.

top_4_nsc_v4 <- top_4_nsc_v4 %>%
  filter(SEASON >= 1997)

# Now let's see how these numbers affect the CB
# as the revenue data is avalible starting from season
# 1996/1997 let's not consider the seasons before it
cor(revenue_gap_cs$TOTAL.REV, top_4_dHHI$HHI)
```

```
## [1] -0.5179731
```

```r
cor(revenue_gap_cs$TOTAL.REV, top_4_HHI$HHI)
```

```
## [1] -0.1122577
```

```r
cor(revenue_gap_cs$TOTAL.REV, C5_top_4$C5)
```

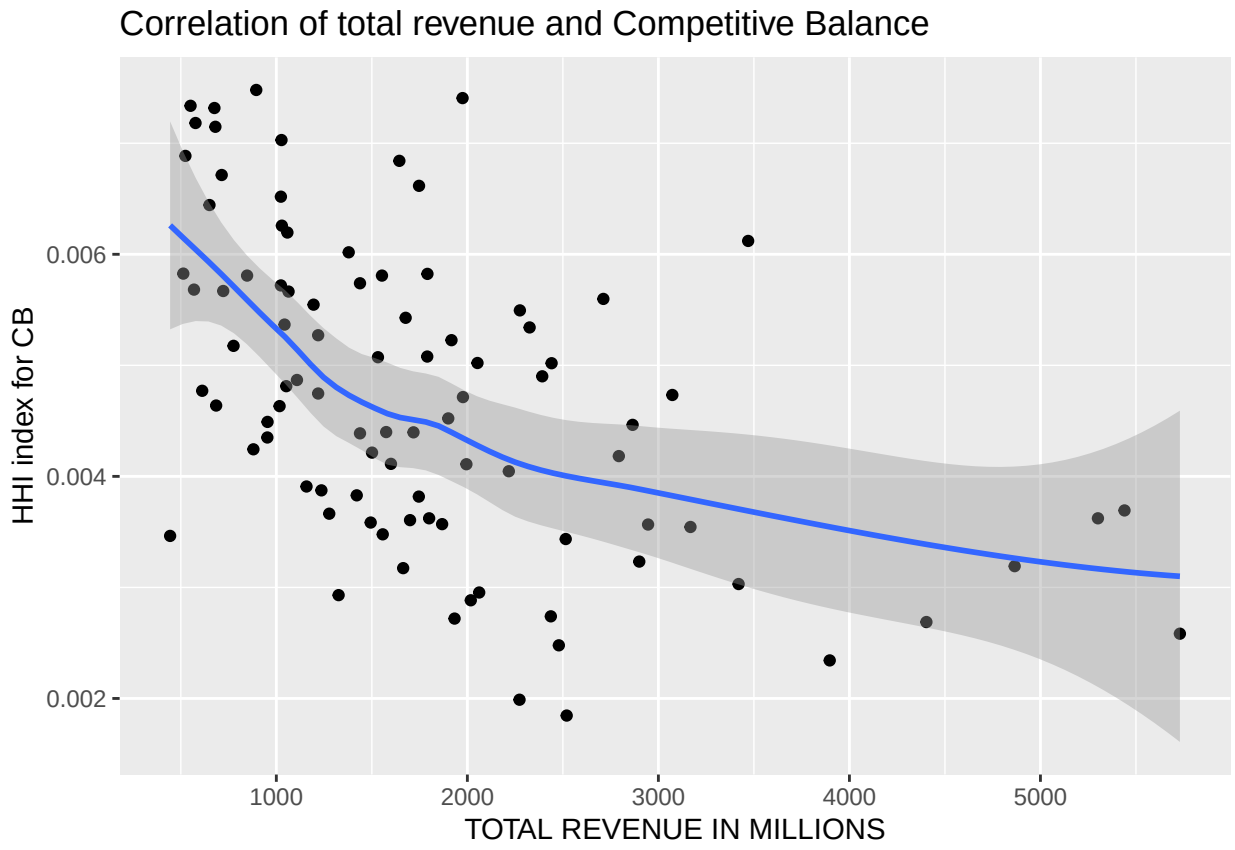
```
## [1] -0.3531471
```

```r
cor(revenue_gap_cs$TOTAL.REV, top_4_nsc_v4$NSN)
```

```
## [1] 0.9280093
```

```r
# All the meausers except the NOLL SKULLY number
# are uncorelated to the total revenue.
# C5 and HHI share similar level of corelation
# with total revenue (-0.35, -0.51), whereas
# HHI holds value of -0.11, being the least correlated
# SO in general higher REVENUE => lower HHI = higher CB
# dHHI is more corelated than HHI as there were many cases
# when the number of teams changed during over seasons,
# and dHHI handles that cases leading to higher corelation.
# For NSN case, the numbers work in opposite direction
# as it has a strong corelation with total revenue, however
# the higher the NSN the lower the CB so the harmony is acquired.

#Now let's plot their corelations

plot_data <- data.frame(SEASON = top_4_dHHI$SEASON, COUNTRY = top_4_dHHI$COUNTRY,
                        TOTAL.REV = revenue_gap_cs$TOTAL.REV, NSN = top_4_nsc_v4$NSN,
                        C5 = C5_top_4$C5, HHI = top_4_HHI$HHI, dHHI = top_4_dHHI$HHI)

plot_data %>%
  ggplot(aes(x = TOTAL.REV, y = dHHI)) +
  geom_point() +
  geom_smooth() +
  labs(x = "TOTAL REVENUE IN MILLIONS", y = "HHI index for CB") +
  ggtitle("Correlation of total revenue and Competitive Balance")
```

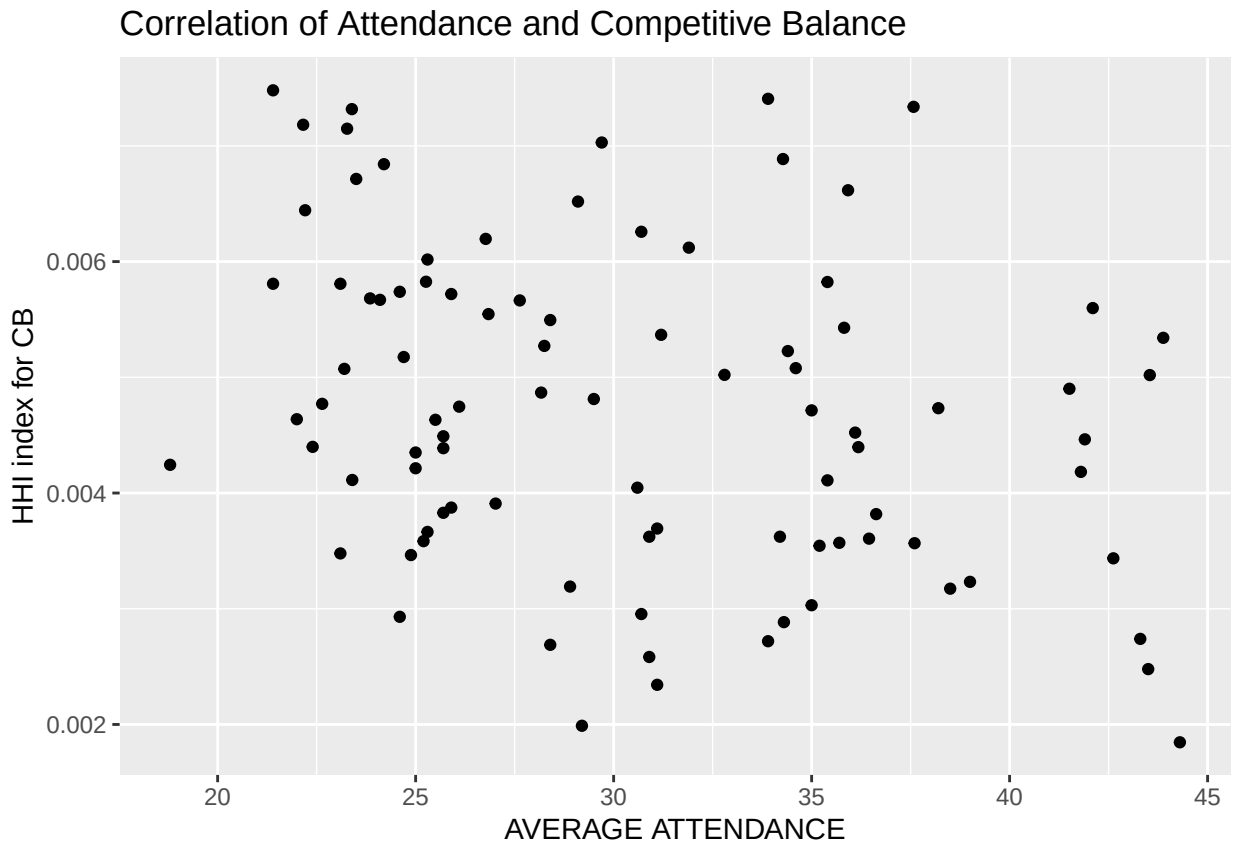
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
# Now let's see how the changes in CB affect the attendance

# Attendance data misses values for 2019 so let's add means by country
att_means <- attendance %>%
  group_by(COUNTRY) %>%
  summarise(MEAN.ATT = mean(ATTENDANCE)) %>%
  mutate(SEASON = "2019")
att_means <- att_means[c(3, 1:2)]
colnames(att_means) <- c("SEASON","COUNTRY","ATTENDANCE")
attendance <- rbind(attendance, att_means)
cor(attendance$ATTENDANCE, top_4_dHHI$HHI)
```

```
## [1] -0.2951501
```

```
# As we can see the corelation is again negative
# for these two factors, let's visualize them for making
# conclusions
ggplot() +
  geom_point(aes(x = attendance$ATTENDANCE, y = top_4_dHHI$HHI)) +
  labs(x = "AVERAGE ATTENDANCE", y = "HHI index for CB") +
  ggtitle("Correlation of Attendance and Competitive Balance")
```




```
# As we can see we have low bias, but high variance
# Seems like the competitive balance value stays stable
# regarding to attendance number but gets higher with the increase
# in attendance in general


# To cap it all, seems like the leading factor is the team revenue
# and the lag factor is the attendance.
# IN GENERAL high REVENUE = high CB = higher ATTENDANCE
# I believe that this connections works on the opposite way two, so
# if we take attendance as a leading factor it would effect the CB
# which in case will effect the leagues revenue.
# Let's check the correlation between attendance and revenue

cor(attendance$ATTENDANCE, revenue_gap_cs$TOTAL.REV)
```

```
## [1] 0.4565764
```

```
# the correlation is positive as excpected

# Dear Mr. Madoyan, I wanted to analyze the investments of the leagues
# in transfermarkt as a leading factor and In my opinion
# the league's ranking would have been the lagging factor, but the data available in the library
# SportsAnalytics270 is very messy and it will take a lot of time for me to
# clean it up. So my leading and lagging variables are attandance(leading) -> revenue(lagging)
```

```
# and revenue(leading) -> revenue(lagging)
# I managed to find a dataset of the transfers from 2000 to 2018 that i can use,
# but at the time of submission I was not able to find a usable data about the rankings of
# the leagues over time.

unique(top_4_standings$LEAGUE)
```

```
## [1] "Premier League"           "Bundesliga 1"
## [3] "La Liga Primera Division" "Serie A"
```

```
unique(transfers$League_to)
```

```
##  [1] Ligue 1                          LaLiga
##  [3] Serie A                          Premier League
##  [5] Super League                     1.Bundesliga
##  [7] Premier Liga                     Liga NOS
##  [9]  Russia                           Scotland
## [11] Championship                     Süper Lig
## [13] Eredivisie                       UAE Gulf League
## [15] Série A                           Qatar
## [17]  United Arab Emirates            Liga MX Clausura
## [19] Liga MX Apertura                 Ligue 2
## [21] MLS                               Portugal
## [23] Torneo Final                     Stars League
## [25] League One                       Primera División
## [27] Segunda División - Segunda Fase  J1 League
## [29] Premiership                      Professional League
## [31] LaLiga2                           Saudi Arabia
## [33] Jupiler Pro League               2.Bundesliga
## [35]  Brazil                          First Division
## [37] Serie B                           Uruguay
## [39] Ledman Liga Pro                   England
## [41]  Mexico                           Argentina
## [43] J1 - 2nd Stage                   Serie C - B
## [45] Bundesliga                       Primavera B
## [47] Superligaen                       China
## [49]  Wales                           J2 League
## [51] SuperLiga                         Bulgaria
## [53]  Venezuela                        Belgium
## [55]  Korea, South                    Eliteserien
## [57]  Denmark                          Croatia
## [59] Allsvenskan                       Sweden
## [61]  Romania                         Second Division (bis 03/04)
## [63]  Czech Republic                   Libya
## [65]  Israel
## 65 Levels: 1.Bundesliga 2.Bundesliga Allsvenskan  Argentina ...  Wales
```

```
leagues <- c("LaLiga", "Serie A", "1.Bundesliga", "Premier League")

top_4_transfers_from <- transfers %>%
  filter(League_from %in% leagues)
```

```r
selling <- top_4_transfers_from %>%
  group_by(Season, League_from) %>%
  summarise(Sold = sum(Transfer_fee))

top_4_transfers_to <- transfers %>%
  filter(League_to %in% leagues)


buying <- top_4_transfers_to %>%
  group_by(Season, League_to) %>%
  summarise(Bought = sum(Transfer_fee))

top_4_dHHI <- top_4_dHHI %>%
  filter(SEASON >= 2001)

buying$Season = top_4_dHHI$SEASON
selling$Season = top_4_dHHI$SEASON

transfer_profits <- data.frame(Season = buying$Season, League = buying$League_to,
                               spent = buying$Bought, earn = selling$Sold)
transfer_profits$profit <- transfer_profits$spent - transfer_profits$earn

# If the profit is lower it means that the country is
# buying more than it is selling, meaning that more and
# more new players join the league, whereas a negative profit
# means that the country's teams mostly sell it's players and does not
# buy alternatives

# Let's again do the checks for correlation
cor(transfer_profits$profit, top_4_dHHI$HHI)
```

```
## [1] 0.1283647
```

```r
# small but positive corellation, meaning that
# the more the profit the higher the dHHI so lower
# the CB.
# Now let's seperate buying and selling cases
cor(transfer_profits$spent, top_4_dHHI$HHI) #0.21
```

```
## [1] 0.2198811
```

```r
cor(transfer_profits$earn, top_4_dHHI$HHI) #0.22
```

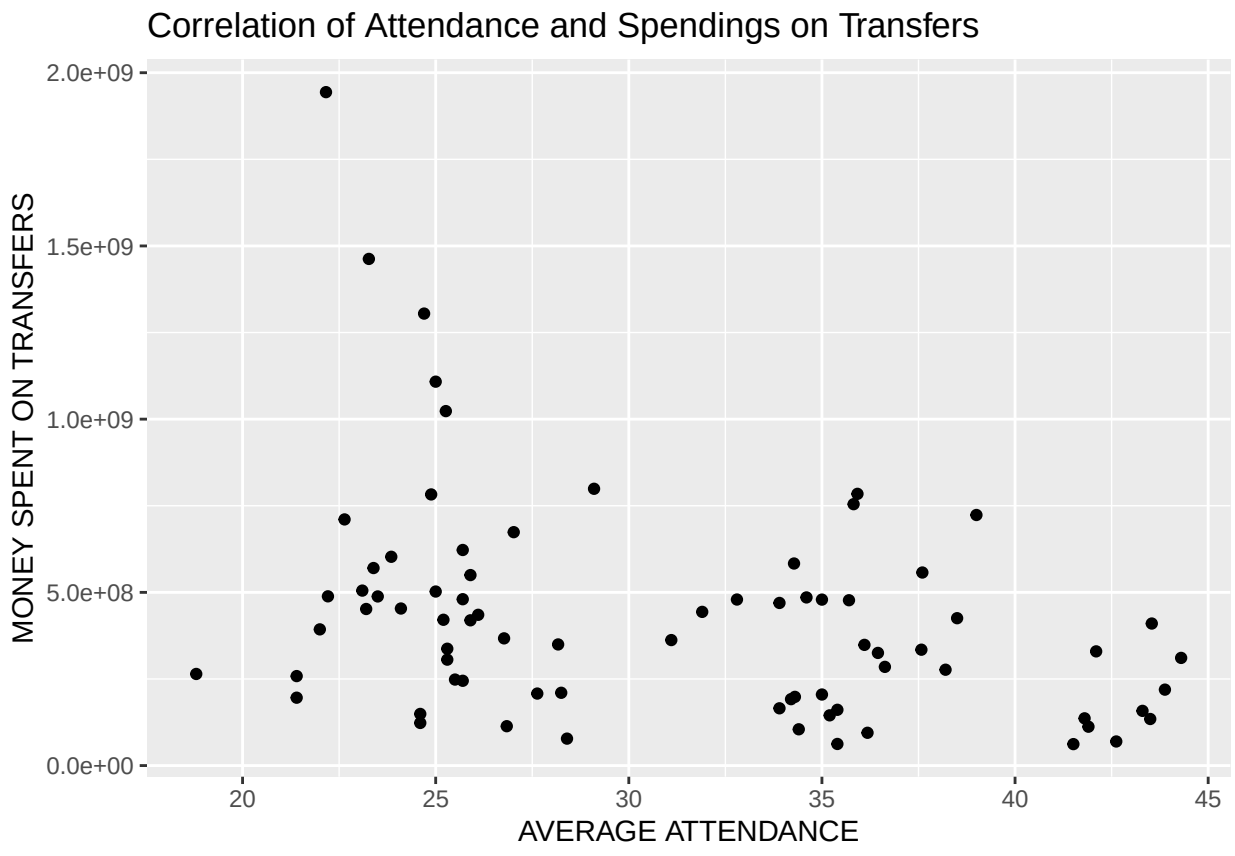
```
## [1] 0.227178
```

```r
# The little difference in correlation in benefit to
# earning. This is somehow connected to the revenew
# analysis, as the money got from selling players
# is part of the leagues revenew.
# However, I believe that the attendacne of the league
# will drop if teams will sell their players
```

```r
# as some funs want to watch the games to see their idols in action

attendance <- attendance %>%
  filter(SEASON > 2000)
cor(transfer_profits$earn, attendance$ATTENDANCE)
```

```
## [1] -0.5374047
```

```r
# The negative correlation supports my opinion

ggplot() +
  geom_point(aes(x = attendance$ATTENDANCE, y = transfer_profits$spent)) +
  labs(x = "AVERAGE ATTENDANCE", y = "MONEY SPENT ON TRANSFERS") +
  ggtitle("Correlation of Attendance and Spendings on Transfers")
```




```r
# We can see that, in the cases of low attendance the leagues spend
# more money on buying new players, and when the attendance is getting higher
# they start to spend less.


# High Profit = High dHHI = low CB = low Attendance
# Profit(leading) Attendance(lagging)
```