EPL ANALYSIS

Dhruvi Shah, Duc Long Ngo & Devansh Rai

BS18bds011, Bj19bds002, BS18bds006

Contents

[Introduction 2](#_Toc40262879)

[Objectives : 2](#_Toc40262880)

[Analysis 3](#_Toc40262881)

[Points per season 3](#_Toc40262882)

[League Position 4](#_Toc40262883)

[Goal Difference 5](#_Toc40262884)

[Home Points v/s Away Points 6](#_Toc40262885)

[Shooting Accuracy Percent 9](#_Toc40262886)

[Linear Regression - total passes v/s total goals 10](#_Toc40262887)

[Gap Between 17th and 18th team 10](#_Toc40262888)

[Goals by headers 11](#_Toc40262889)

[Goals by freekicks 12](#_Toc40262890)

[Total passes over different seasons 12](#_Toc40262891)

[Total tackles over different seasons 13](#_Toc40262892)

[Total passes v/s total tackles over different seasons 13](#_Toc40262893)

[Total through balls over different seasons 14](#_Toc40262894)

[APPENDIX 15](#_Toc40262895)

# Introduction

English Premier League is a football league comprising of 20 teams per season. Each team plays 38 matches per season. For every match a team wins the team receives 3 points and for every match it draws, the team receives 1 point. Football is an extremely competitive sport and thus it is essential to analyse the changes that have taken place in this sport to be able to make future predictions.

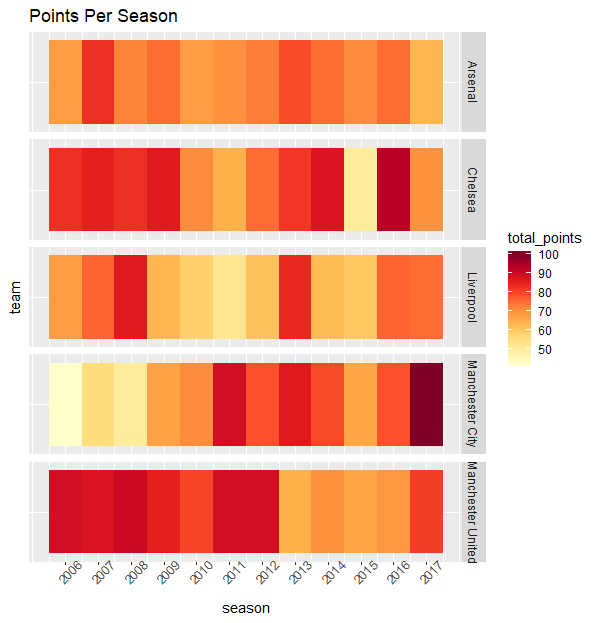
Dataset variables description : https://www.kaggle.com/zaeemnalla/premier-league#results.csv

## Objectives :

1. The report aims at analysing how football in the EPL has changed over the years.
2. Comparing winning rate between HOME and AWAY ground over the year.
3. Compare goals scored by header and feet over the year.
4. Analyse change in league positions for the top 5 teams
5. Calculate shooting Accuracy
6. Find the points difference between teams that remain in the EPL and teams that get relegated

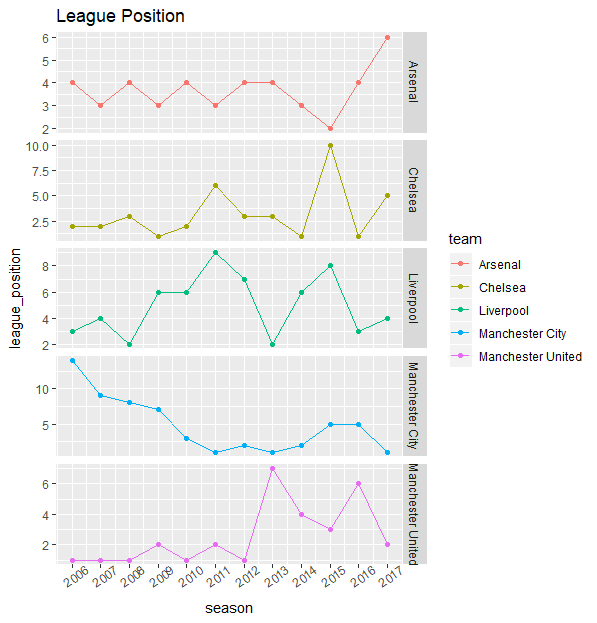
# Analysis

## Points per season



Manchester city shows the most overall improvement out of all the teams, the club started with less than 50 points and ended up scoring the overall maximum points that is 100 points in the 2017-2018 season. On the other hand City’s biggest rivals Manchester United had an incredible start but ended up deteriorating the most. Manchester City improved under Pep Guaridola who’s tenure started from 2016 and he was able to get them their 2017 title with a significant margin. Sir Alex Ferguson was the manager for Manchester United till 2013 May, and the club did really well under him but when the signed David Moyes as their manager the club took a very big hit.

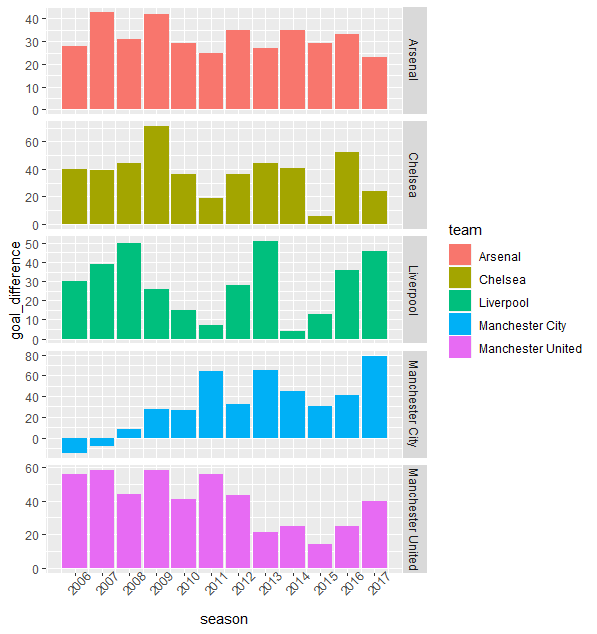
## League Position



Under several mangers where United saw a continuous downfall after Ferguson left. Arsenal went from an average 4th position to the 6th position. Even Chelsea started off from at position 2 and ended up at position 5.

* Arsenal had shown the sustainability from 2006-2007 to 2015-2016 when Arsene Wenger was still the manager of the club and lost their position after he left.
* Chelsea constantly shows the fluctuation in position since 2006-2007 season because of various changes in coaches. But they still show their strength and won the title in season 2009-2010, 2014-2015, 2016-2017.
* Liverpool shows the most number of inconsistencies, from the 2011-2012 season they did not do so well except for the season 2013-2014 when they almost won the title if it hadn’t been for Gerald’s slip.
* City has shown the most improvement over the years. Since the new boss came into power, they has been injected good amount of money to reconstruct the team. They won the title in season 2011-2012 with the same points as Manchester United, 2013-2014, 2017-2018. Manchester City has become the new rising power of the EPL.
* Manchester United shows a consistence in Sir Alex Ferguson still in power. They won 3 titles in a row from 2006-2007 to 2008-2009. After Ferguson left, they struggled to get back their dominance in the league.

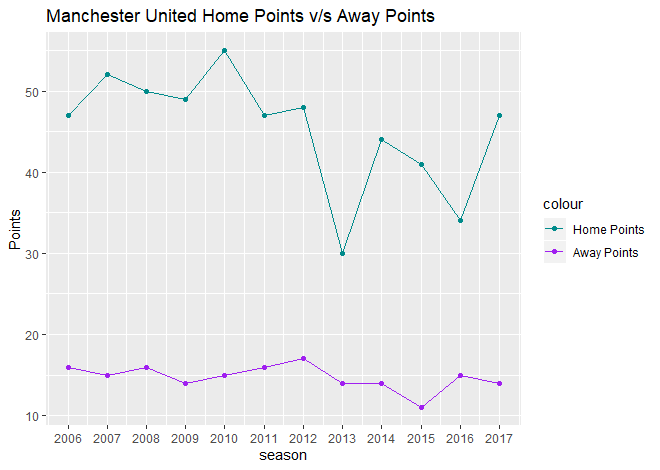
## Goal Difference

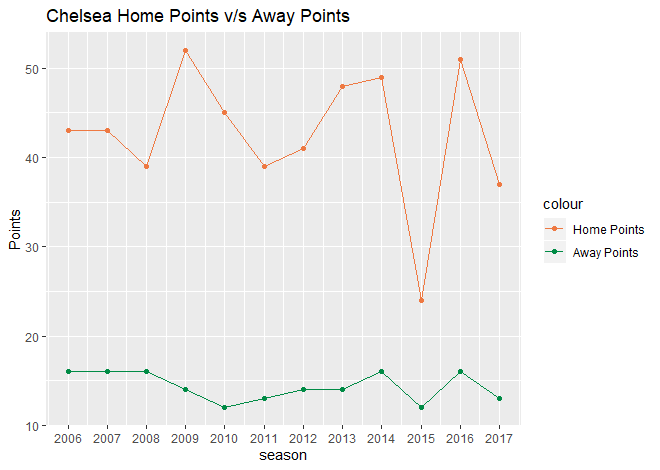


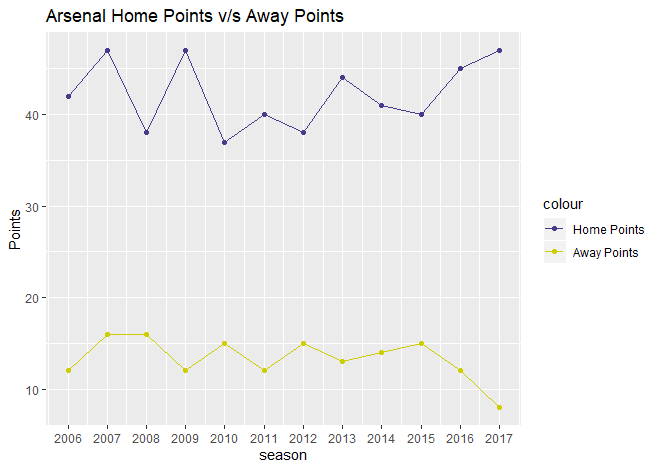
Goal difference is a very important factor in football when two teams end up with the same points at the end of the league, they check the number of goals conceded by that team to make a decision. Higher the number of goals conceded shows a weak defense strategy for that team.

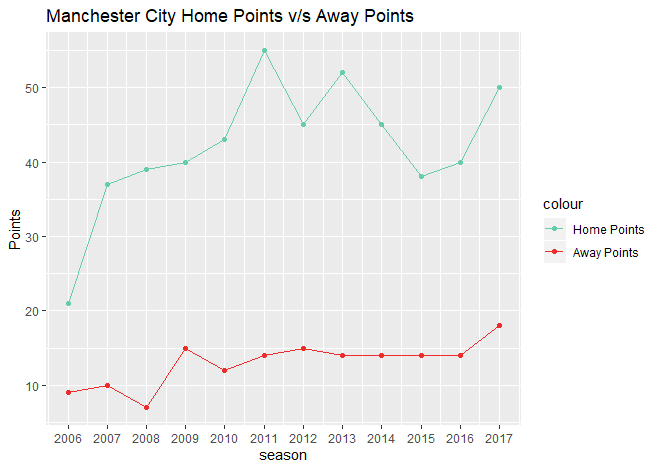
* Arsenal has shown consistency by keeping their goal difference at the level of 30-40 for 11 consecutive years
* Chelsea has good and bad seasons. Their best season was in 2009-2010, their goal difference was more than 60, highest in the EPL in this season. Besides, they are also known to be practical in football, they doesn’t have to score much but still maximize the points they got and won the title in the year of 2014-2015 and 2016-2017 without having an outstanding goal difference compared to other teams.
* Liverpool was so close to get the title in season 2008-2009, they got the dominant goal difference but still lost the title in the hand of Manchester United. There are a lot of season when they show their dominance in goal differences but still can not maximize their points gained and lost the title to other teams.
* City got the highest goal difference in season 2017-2018 with nearly 80 goal difference, leaving behind all of other teams in the race to the title.
* MU was very effective. In season 2008-2009, they have less goal difference compared to Liverpool but still manage to win both the EPL and Champions League after an intense final against Chelsea.

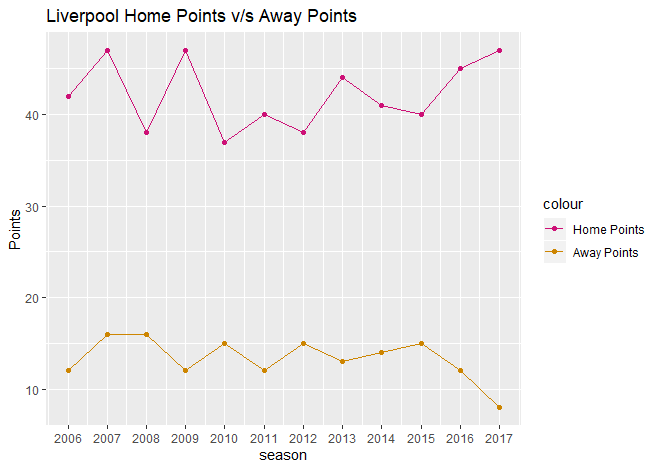
## Home Points v/s Away Points







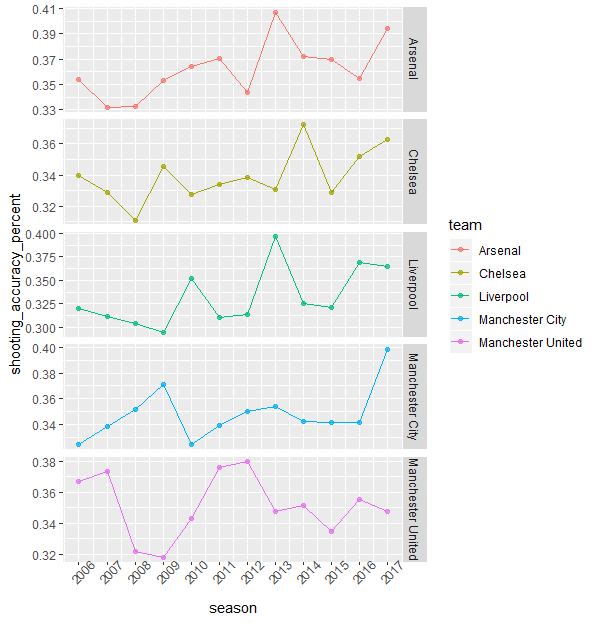




These graphs show that even the top teams do not manage to score enough goals at their away grounds, and that playing at home grounds acts in favour of most teams. This home advantage is due to the familiarity with the ground and having their fans there. Comparing this graph to the points scored shows teams that score a lot of home points but less number of away points do not do as well as teams scoring higher away points.

From these 5 graphs, we can see the significant differences in the performance of top 5 team in Home ground vs Away ground. Teams that have the potential to do better should focus on away games as a chance to again additional points. Managers should play their A teams on away grounds because each away win is extremely essential.

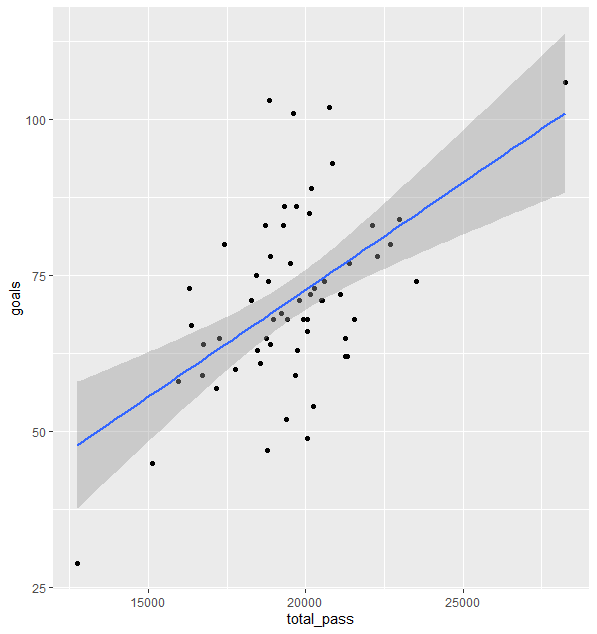
## Shooting Accuracy Percent



Shooting accuracy is calculated using ontarget\_scoring\_att (that is shots on target) divided by the total\_scoring\_att (that is total shots)

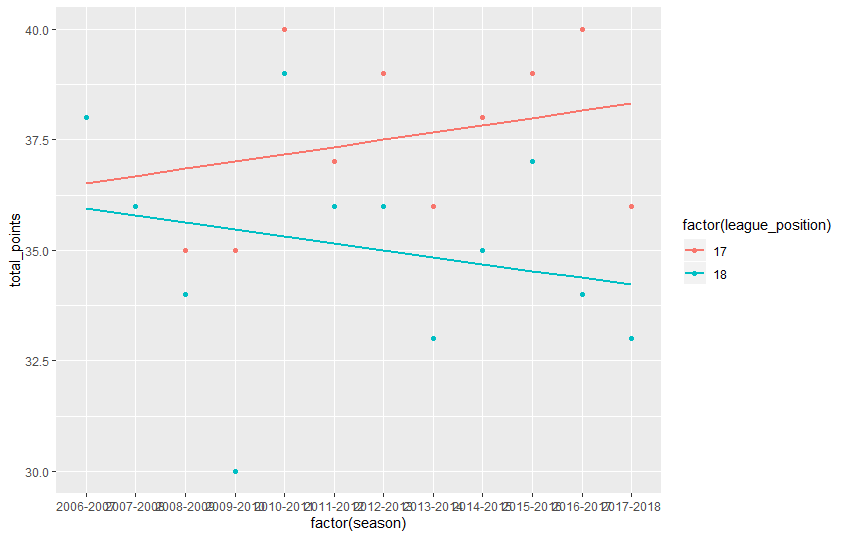
Despite having the most number of points(total of 12 seasons) Manchester United has a shooting accuracy much lesser than the other teams. This means that the make multiple attempts at shooting but only a few are actually on target. This is quite unexpected as one would expect a top scoring club to have the best statistics. We can also say that United have play a more attacking game. On the other hand Manchester City has shown improvement in its shooting accuracy. Liverpool has been extremely inconsistent and so have Chelsea and Arsenal, although of the top teams Arsenal has had the best shooting accuracy in the year 2013-2014.

## Linear Regression - total passes v/s total goals



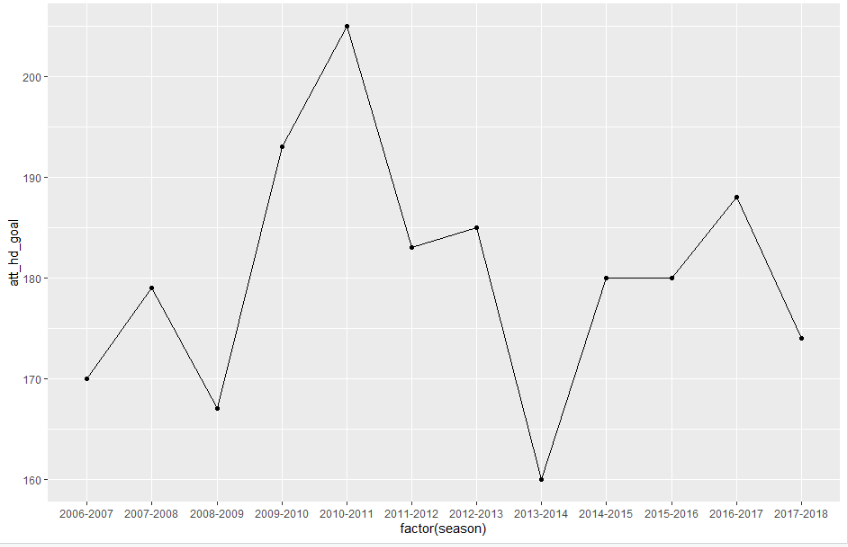
There is a correlation between the total passes and the total goals, more passes means more possession of the ball with the team which is good because this can create more opportunities to score goals. Having more possession of the ball is a key factor and can often change the chances of a team winning or losing.

## Gap Between 17th and 18th team



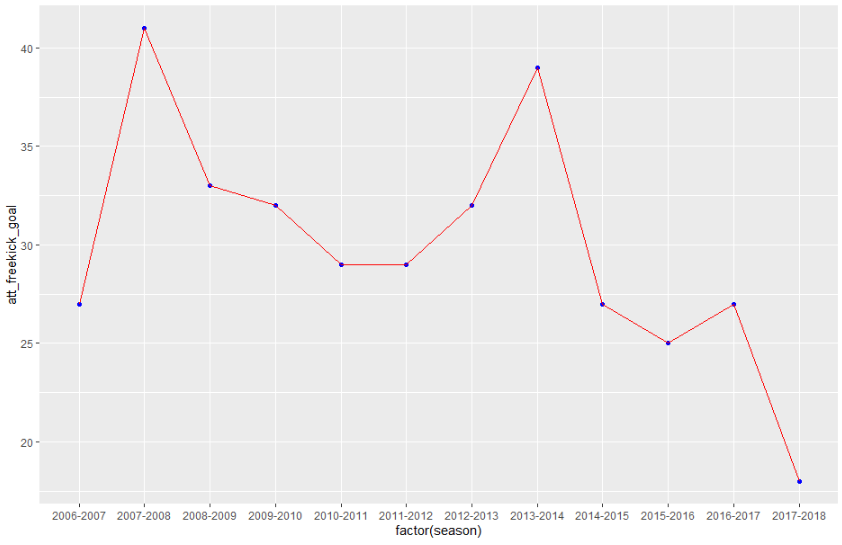
From the graph, we can see that as the year goes by, the gap between the 17-th team and 18-th team, between the team which can stay and will be relegated is getting wider. We also can say that here is a bigger differences in performances of the team who stays versus the one who will be relegated

## Goals by headers



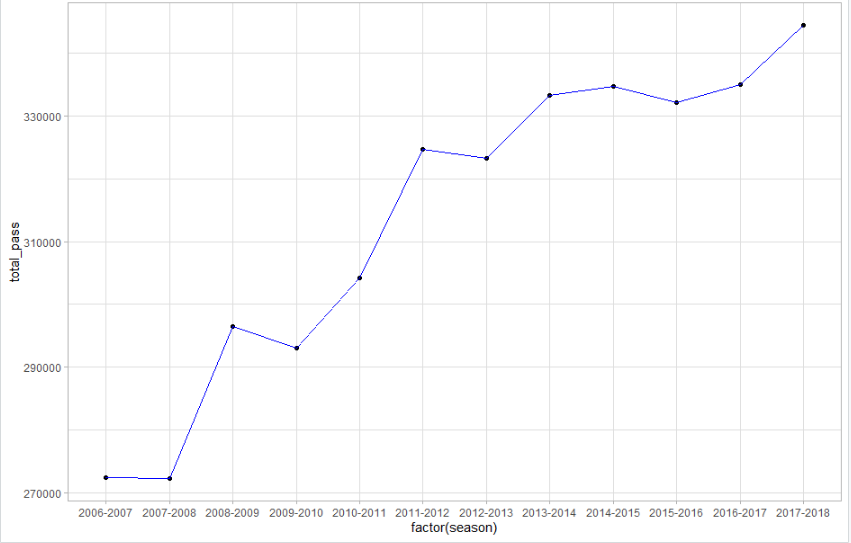
Scoring by header was a trend from season 2009-2010 to season 2012-2013. These is also the time when header becomes the strength of English football. But later on, different game approach has been introduced and header showed to be less effective.

## Goals by freekicks



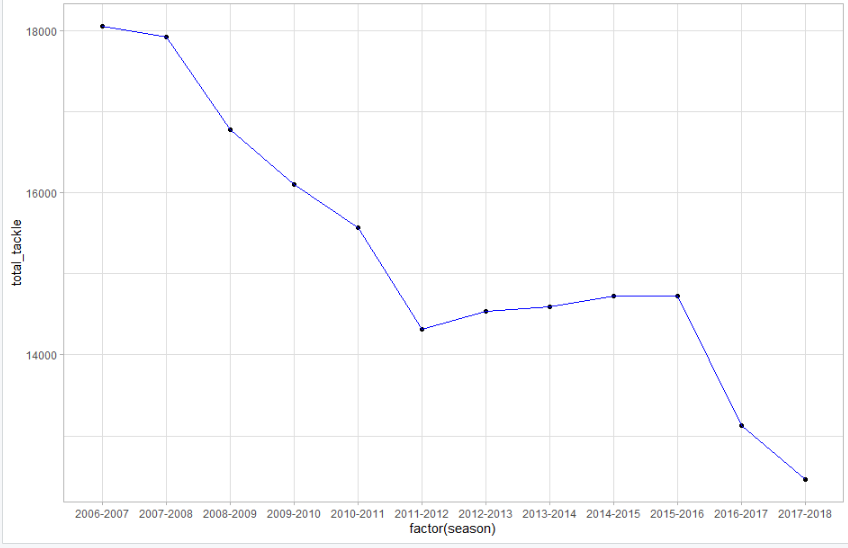
Number of goals scored from freekicks decrease heavily. The number of goal scored by freekick peaked in 2007-2008 season and felt back to around 30 to 35 in 5 continuous years after peaked again in 2013-2014 season and droped heavily after that.

## Total passes over different seasons



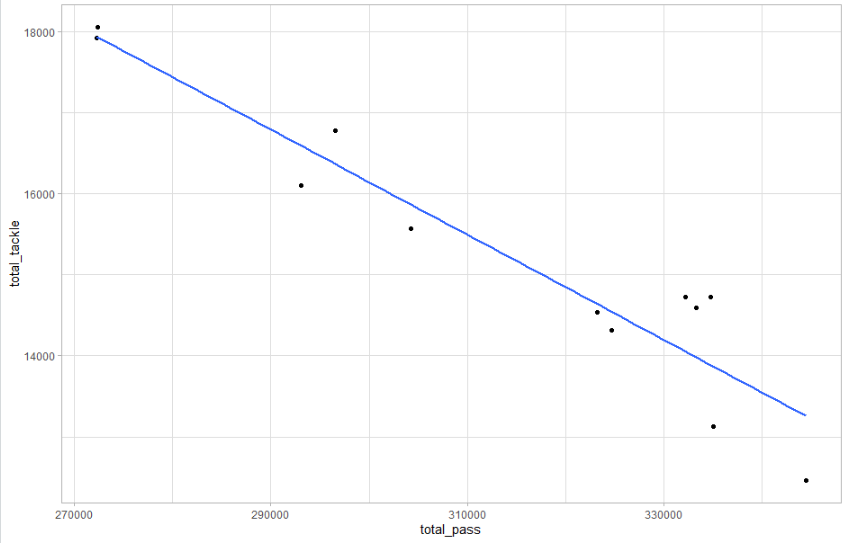
There is a huge increase in number of pass over the years. The total pass in season 2006-2007 was nearly 27000 but increase by 50% in 11 years. These can show the big changes in English football. Modern football focus now is shifted to short passes and possess the ball.

## Total tackles over different seasons



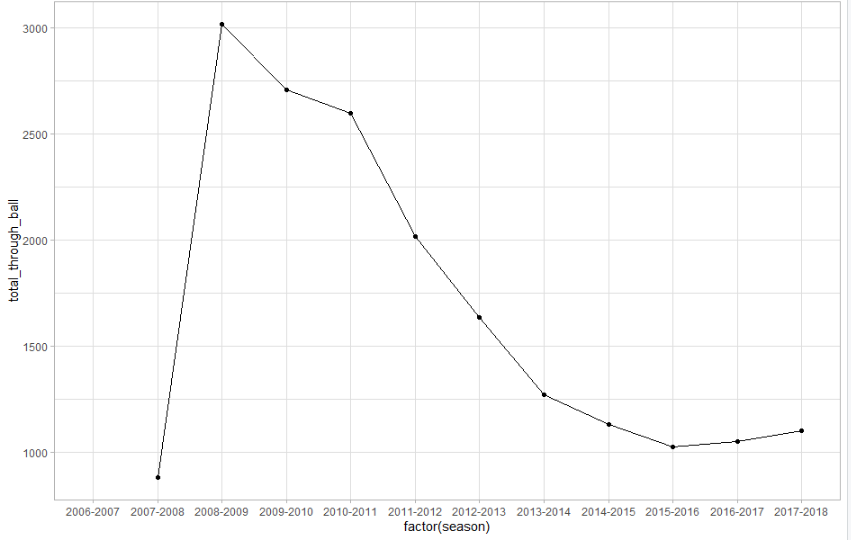
As the number of pass increase, the number of tackle decrease. It is more difficult to get the ball when the team play tiki taka and make a lot of passes and possess the ball. That’s also why modern football focus shifted to ball possession and short pass.

## Total passes v/s total tackles over different seasons



The graph below shows the correlation between number of pass and number of tackle. As the graph below shows, there is a negative correlation between total pass and total tackle.

## Total through balls over different seasons



Total of through ball peaked in season 2008-2009 with 3000 tackles and declined slowly in 7 continuous seasons. It reduced 3-times in 10 years from 3000 tackles to just more that 1000 in season 2017-2018. These changes happened due to the increase in number of short passes. In modern football, short passes is the key to success. Long passes and through ball is reduced and only being used to surprise the opponents. Constantly using through ball is considered a waste of ball in modern football.

# APPENDIX

**#importing packages**

library(dplyr)

library(RColorBrewer)

library(ggplot2)

**#Import the dataset and select suitable columns for analysis**

epl <- read.csv("stats.csv")

epl <- epl[, c(1,2,3,4,7,8,17,18,21,22,29,42)]

**#Add points per season and league position column**

epl <- epl %>% mutate(total\_points = (wins\*3)+(38-(wins + losses)))

season\_index=seq(6,17)

ffillv <- function(i) rep(season\_index[i], 20)

season\_index<-c(sapply(seq\_len(length(season\_index)), ffillv))

epl<-epl%>%mutate(season\_index=season\_index)

epl<-epl[order(epl['season\_index'],-epl['total\_points']),]

epl<- epl %>% mutate(league\_position=rep(seq(1,20,1), times = 12))

## **Compute a dataset to find the top 5 teams in EPL**

**#creates a list of unique team names**

unique\_teams <- unique(epl$team)

**#creates an empty vector**

total\_wins <- c()

**#creates a for loop that returns a vector containing the sum of wins for each team over all the seasons**

for (i in 1:length(unique\_teams)){

team <- epl[epl$team == unique\_teams[i],]

total\_wins[i] = sum(team$wins)}

#creates an empty vector

total\_losses <- c()

**#creates a for loop that returns a vector containing the sum of losses for each team over all the seasons**

for (i in 1:length(unique\_teams)){

team <- epl[epl$team == unique\_teams[i],]

total\_losses[i] <- sum(team$losses)}

### **#creates a dataframe with 3 columns containing total wins, losses for each team**

points\_table <- data.frame(team = unique\_teams, total\_wins, total\_losses)

**#creates a new column total\_points that contains the total points over 12 seasons**

points\_table <- points\_table %>% mutate(total\_points = (total\_wins\*3) + ((38\*12) - (total\_wins + total\_losses)))

After computing the dataset to keep records of all the teams’ total point from season 2006-2007 to season 2017-2018, we arrange it in descending order with respect to total points and then filter the top 5 teams.

**#arranges total\_points in a descending order which gives us the top 5 and bottom 5 teams to analyse**

points\_table <- points\_table %>% arrange(desc(total\_points))

**#creates a vector of top 5 team names**

top\_teams <- points\_table$team[1:5]

**#creates a dataset with the top 5 teams**

epl\_top\_teams <- epl %>% filter(team %in% top\_teams)

**#creating a colour palate**

myColors <- brewer.pal(9, "YlOrRd")

**#plotting season wise points for each team**

ggplot(epl\_top\_teams, aes(season, team, fill = total\_points)) + geom\_tile() + facet\_grid(team~., scales = "free\_y") + scale\_x\_continuous(breaks = seq(2006, 2017)) +scale\_fill\_gradientn(colors = myColors) + theme(axis.text.y = element\_blank(), axis.ticks.y = element\_blank(), axis.text.x = element\_text(angle = 45)) + labs(title = "Points Per Season")

**#plotting changes in league position over the seasons**

ggplot(epl\_top\_teams, aes(season, league\_position, col = team)) + geom\_line() + geom\_point() + facet\_grid(team~., scales = "free\_y") + scale\_x\_continuous(breaks = seq(2006,2017)) + theme(axis.text.x = element\_text(angle = 35))

**#calculating goal difference**

epl\_top\_teams <- epl\_top\_teams %>% mutate(goal\_difference = goals - goals\_conceded)

**#plotting season against goal difference for that season for the top 5 teams**

ggplot(epl\_top\_teams, aes(season, goal\_difference, fill = team)) + geom\_bar(stat = "identity") + facet\_grid(team~., scales = "free\_y") + scale\_x\_continuous(breaks = seq(2006,2017)) + theme(axis.text.x = element\_text(angle = 45))

**#calculating home points for the top 5 teams for each season**

results<- read.csv(‘result.csv’)

home\_points <- results %>% filter(home\_team %in% top\_teams) %>% filter(result %in% c('H', 'D'))

home\_points <- home\_points [, c(1,3,5,6)]

home\_points <- home\_points %>% mutate(game\_points = 3)

for (i in 1:length(home\_points$results)){

if (home\_points$results[i] == ‘D’){

home\_points$game\_points = 1}}

**#creating empty vector to store home points of top 5 teams**

man\_u\_home <- c()

che\_home <- c()

ars\_home <- c()

man\_c\_home <- c()

liv\_home <- c()

**# fill all empty vector with points of home ground for each season using a for loop**

**Manchester United**

for (i in 1:length(unique\_seasons)){

x <- home\_points %>% filter(home\_team == 'Manchester United') %>% filter(season == unique\_seasons[i])

man\_c\_home[i] <- sum(x$game\_points)}

**Chelsea**

for (i in 1:length(unique\_seasons)){

x <- home\_points %>% filter(home\_team == 'Chelsea') %>% filter(season == unique\_seasons[i])

che\_home[i] <- sum(x$game\_points)}

**Arsenal**

for (i in 1:length(unique\_seasons)){

x <- home\_points %>% filter(home\_team == 'Arsenal') %>% filter(season == unique\_seasons[i])

ars\_home[i] <- sum(x$game\_points)}

**Manchester City**

for (i in 1:length(unique\_seasons)){

x <- home\_points %>% filter(home\_team == 'Manchester City') %>% filter(season == unique\_seasons[i])

man\_c\_home[i] <- sum(x$game\_points)}

**Liverpool**

for (i in 1:length(unique\_seasons)){

x <- home\_points %>% filter(home\_team == 'Liverpool') %>% filter(season == unique\_seasons[i])

liv\_home[i] <- sum(x$game\_points)}

**#creating a dataframe containing the home points of each team for every season**

points\_at\_home <- data.frame(man\_u\_home, che\_home, ars\_home, man\_c\_home, liv\_home)

**#calculating away points for the top 5 teams for each season**

away\_points <- results %>% filter(away\_team %in% top\_teams) %>% filter(result %in% c('A', 'D'))

away <- away\_points[, c(2,4,5,6)]

away <- away %>% mutate(game\_points = 3)

for (i in 1:length(away$result)){

if (away$result[i] == ‘D’){

away$game\_points = 1}}

**#creating empty vector to store away points of top 5 teams**

man\_u\_away <- c()

che\_away <- c()

ars\_away <- c()

man\_c\_away <- c()

liv\_away <- c()

**# fill all empty vectors with points of away ground for each season using a for loop**

**Manchester United**

for (i in 1:length(unique\_seasons)){

x <- away %>% filter(away\_team == 'Manchester United') %>% filter(season == unique\_seasons[i])

man\_u\_away[i] <- sum(x$game\_points)}

**Chelsea**

for (i in 1:length(unique\_seasons)){

x <- away%>% filter(away\_team == 'Chelsea') %>% filter(season == unique\_seasons[i])

che\_away [i] <- sum(x$game\_points)}

**Arsenal**

for (i in 1:length(unique\_seasons)){

x <- away %>% filter(away\_team == 'Arsenal') %>% filter(season == unique\_seasons[i])

ars\_away [i] <- sum(x$game\_points)}

**Manchester City**

for (i in 1:length(unique\_seasons)){

x <- away %>% filter(away\_team == 'Manchester City') %>% filter(season == unique\_seasons[i])

man\_c\_away[i] <- sum(x$game\_points)}

**Liverpool**

for (i in 1:length(unique\_seasons)){

x <- away %>% filter(away\_team == 'Liverpool') %>% filter(season == unique\_seasons[i])

liv\_away[i] <- sum(x$game\_points)}

**#creating a dataframe containing the away points of each team for every season**

points\_at\_away <- data.frame(man\_u\_away, che\_away, ars\_away, man\_c\_away, liv\_away)

**#adding points per season to a team specific vector which is later compiled into one big data frame**

mu\_total\_pts <- epl\_top\_teams[epl\_top\_teams$team == "Manchester United", "total\_points"]

che\_total\_pts <- epl\_top\_teams[epl\_top\_teams$team == "Chelsea", "total\_points"]

ars\_total\_pts <- epl\_top\_teams[epl\_top\_teams$team == "Arsenal", "total\_points"]

mc\_total\_pts <- epl\_top\_teams[epl\_top\_teams$team == "Manchester City", "total\_points"]

liv\_total\_pts <- epl\_top\_teams[epl\_top\_teams$team == "Liverpool", "total\_points"]

**#combing home points and away points into one dataframe**

points\_at\_home\_away <- cbind(points\_at\_home, points\_at\_away)

points\_at\_home\_away <- points\_at\_home\_away %>% mutate(mu\_tp = mu\_total\_pts)%>% mutate(che\_tp = che\_total\_pts) %>% mutate(ars\_tp = ars\_total\_pts) %>% mutate(mc\_tp = mc\_total\_pts) %>% mutate(liv\_tp = liv\_total\_pts) %>% mutate(season = unique(epl\_top\_teams$season))

**#plotting home v/s away points for Manchester United**

ggplot(points\_at\_home\_away, aes(x = season)) + geom\_line(aes(y = man\_u\_home, col = "cyan4")) + geom\_point (aes(y = man\_u\_home, col = "cyan4")) + geom\_line(aes(y = man\_u\_away, col = "purple")) + geom\_point (aes(y = man\_u\_away, col = "purple")) + scale\_color\_manual(labels = c("Home Points", "Away Points"), values = c("cyan4", "purple")) + scale\_x\_continuous(breaks = seq(2006,2017)) + labs(title = "Manchester United Home Points v/s Away Points", y = "Points")

**#plotting home v/s away points for Chelsea**

ggplot(points\_at\_home\_away, aes(x = season)) + geom\_line(aes(y = che\_home, col = "sienna2")) + geom\_point (aes(y = che\_home, col = "sienna2")) + geom\_line(aes(y = che\_away, col = "springgreen4")) + geom\_point (aes(y = che\_away, col = "springgreen4")) + scale\_color\_manual(labels = c("Home Points", "Away Points"), values = c("sienna2", "springgreen4")) + scale\_x\_continuous(breaks = seq(2006,2017)) + labs(title = "Chelsea Home Points v/s Away Points", y = "Points")

**#plotting home v/s away points for Arsenal**

ggplot(points\_at\_home\_away, aes(x = season)) + geom\_line(aes(y = ars\_home, col = "slateblue4")) + geom\_point (aes(y = ars\_home, col = "slateblue4")) + geom\_line(aes(y = ars\_away, col = "yellow3")) + geom\_point (aes(y = ars\_away, col = "yellow3")) + scale\_color\_manual(labels = c("Home Points", "Away Points"), values = c("slateblue4", "yellow3")) + scale\_x\_continuous(breaks = seq(2006,2017)) + labs(title = "Arsenal Home Points v/s Away Points", y = "Points")

**#plotting home v/s away points for Manchester City**

ggplot(points\_at\_home\_away, aes(x = season)) + geom\_line(aes(y = man\_c\_home, col = "aquamarine3")) + geom\_point (aes(y = man\_c\_home, col = "aquamarine3")) + geom\_line(aes(y = man\_c\_away, col = "firebrick2")) + geom\_point (aes(y = man\_c\_away, col = "firebrick2")) + scale\_color\_manual(labels = c("Home Points", "Away Points"), values = c("aquamarine3", "firebrick2")) + scale\_x\_continuous(breaks = seq(2006,2017)) + labs(title = "Manchester City Home Points v/s Away Points", y = "Points")

**#plotting home v/s away points for Liverpool**

ggplot(points\_at\_home\_away, aes(x = season)) + geom\_line(aes(y = ars\_home, col = "deeppink3")) + geom\_point (aes(y = ars\_home, col = "deeppink3")) + geom\_line(aes(y = ars\_away, col = "orange3")) + geom\_point (aes(y = ars\_away, col = "orange3")) + scale\_color\_manual(labels = c("Home Points", "Away Points"), values = c("deeppink3", "orange3")) + scale\_x\_continuous(breaks = seq(2006,2017)) + labs(title = "Liverpool Home Points v/s Away Points", y = "Points")

**#calculating shooting accuracy**

epl\_top\_teams <- epl\_top\_teams %>% mutate(shooting\_accuracy\_percent = ontarget\_scoring\_att/total\_scoring\_att)

#**#plotting shooting accuracy for each team**

ggplot(epl\_top\_teams, aes(season, shooting\_accuracy\_percent, col = team)) +geom\_point(alpha = 0.7) + geom\_line() + scale\_x\_continuous(breaks = seq(2006, 2017)) + theme(axis.text.x = element\_text(angle = 45)) + facet\_grid(team~., scales = "free\_y")

**#Import the dataset and compute the columns for number of draws, total points and in\_season\_ranking (same as above)**

**#Select suitable columns for analysis:**

grouped\_df<- df%>% group\_by(season)%>% summarise(total\_scoring\_att = sum(total\_scoring\_att),ontarget\_scoring\_att=sum(ontarget\_scoring\_att),att\_hd\_goal=sum(att\_hd\_goal),att\_freekick\_goal=sum(att\_freekick\_goal),att\_ibox\_goal=sum(att\_ibox\_goal),att\_obox\_goal=sum(att\_obox\_goal),clean\_sheet=sum(clean\_sheet),goals\_conceded=sum(goals\_conceded),total\_tackle=sum(total\_tackle),total\_pass=sum(total\_pass),total\_through\_ball=sum(total\_through\_ball),total\_yel\_card=sum(total\_yel\_card),total\_red\_card=sum(total\_red\_card))

**#plotting total goal scored by header in each season**

ggplot(grouped\_df,aes(x=factor(season),y=att\_hd\_goal,group=1))+geom\_point()+geom\_line()

**#plotting total of goals scored by freekick**

ggplot(grouped\_df,aes(x=factor(season),y=att\_freekick\_goal,group=1))+geom\_point(col='blue')+geom\_line(col='red')

**#plotting total of pass in each season**

ggplot(grouped\_df,aes(x=factor(season),y=total\_pass,group=1))+geom\_point()+geom\_line(col='blue')+theme\_light()

**#plotting total of tackle in each season:**

ggplot(grouped\_df,aes(x=factor(season),y=total\_tackle,group=1))+geom\_point()+geom\_line(col='blue')+theme\_light()

**#plotting linear regression model for total\_pass v/s total\_tackle**

ggplot(grouped\_df,aes(x=total\_pass,y=total\_tackle))+geom\_point()+geom\_smooth(method='lm',se=FALSE)+theme\_light()

**#plotting total of through ball in each season:**

ggplot(grouped\_df,aes(x=factor(season),y=total\_through\_ball,group=1))+geom\_point()+ geom\_line()+theme\_light()