Chelsea_FC_R_Code

Akshay Krishnan

2022-10-10

Part A

Problem 1

```
# Reading the necessary packages
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(stringr)
```

Read the dataset from the local folder

Dataset Source - Link

```
#Importing the dataset in R
df <- read_csv("../Datasets/PL_Results.csv")
```

```
## Rows: 11037 Columns: 23
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): Season, HomeTeam, AwayTeam, FTR, HTR, Referee
## dbl  (16): FTHG, FTAG, HTHG, HTAG, HS, AS, HST, AST, HC, AC, HF, AF, HY, AY,...
## dttm  (1): DateTime
##
## 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.
```

```
#Printing the first few lines of the dataset
head(df)
```

```
## # A tibble: 6 x 23
##   Season  DateTime            HomeTeam    AwayTeam  FTHG  FTAG FTR    HTHG  HTAG
##   <chr>   <dttm>              <chr>       <chr>    <dbl> <dbl> <chr> <dbl> <dbl>
## 1 1993-94 1993-08-14 00:00:00 Arsenal     Coventry     0     3 A        NA    NA
## 2 1993-94 1993-08-14 00:00:00 Aston Villa QPR          4     1 H        NA    NA
## 3 1993-94 1993-08-14 00:00:00 Chelsea     Blackbu~     1     2 A        NA    NA
## 4 1993-94 1993-08-14 00:00:00 Liverpool   Sheffie~     2     0 H        NA    NA
```

```
## 5 1993-94 1993-08-14 00:00:00 Man City    Leeds         1     1 D         NA    NA
## 6 1993-94 1993-08-14 00:00:00 Newcastle   Tottenh~      0     1 A         NA    NA
## # ... with 14 more variables: HTR <chr>, Referee <chr>, HS <dbl>, AS <dbl>,
## #   HST <dbl>, AST <dbl>, HC <dbl>, AC <dbl>, HF <dbl>, AF <dbl>, HY <dbl>,
## #   AY <dbl>, HR <dbl>, AR <dbl>
```

```
# Checking the number of rows and columns
print(paste("There are ",
            nrow(df),
            " rows in the dataset and ",
            ncol(df),
            " columns in the dataset"))
```

```
## [1] "There are  11037  rows in the dataset and  23  columns in the dataset"
```

Dataset Description

The dataset contains **11,037 rows** and **23 columns** describing all of the Premier League Fixtures, as well as the results across all seasons(1993-94 to 2021-22)

Description of the Variables

Column Name	Variable Explanation
Season	Season Year
DateTime	Match Date and Time (yyyy-mm-dd hh:mm:ss)
HomeTeam	Home Team
AwayTeam	Away Team
FTHG	Full Time Home Team Goals
FTAG	Full Time Away Team Goals
FTR	Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG	Half Time Home Team Goals
HTAG	Half Time Away Team Goals
HTR	Half Time Result (H=Home Win, D=Draw, A=Away Win)
Referee	Match Referee
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HC	Home Team Corners
AC	Away Team Corners
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HY	Home Team Yellow Cards
AY	Away Team Yellow Cards
HR	Home Team Red Cards
AR	Away Team Red Cards

Dataset Tidying Process

Creating a final table(tidy dataset) from the source dataset. This tidy dataset would contain the final points, goal tally and league position for each team at the end of every Season of the Premier League

Steps to create the tidy table

1. Creating new variables such as the final result and points earned for the Home Team and Away Team in each game from the source dataset

```
#Creating variables such as Home Team Points Earned, Away Team Points Earned
#and also Flag variables showing the Wins,Losses and Draws for the Home and Away Team
df <- df %>%
  mutate(
    HT_Pts_Earned = case_when(FTR == 'H' ~ 3, FTR == 'A' ~ 0, FTR == 'D' ~ 1),
    AT_Pts_Earned = case_when(FTR == 'H' ~ 0, FTR == 'A' ~ 3, FTR == 'D' ~ 1),
    HT_Win  = ifelse(FTR == 'H' , 1, 0),
    HT_Draw = ifelse(FTR == 'D' , 1, 0),
    HT_Loss = ifelse(FTR == 'A' , 1, 0),
    AT_Win  = ifelse(FTR == 'A' , 1, 0),
    AT_Draw = ifelse(FTR == 'D' , 1, 0),
    AT_Loss = ifelse(FTR == 'H' , 1, 0))
```

2. Grouping the seasonwise results of the Home Team and the Away Team and then combining both those tables to create a finalised league table containing the Season, Team Name, Points earned, Goals Scored, Goals Conceeded,Wins, Draws and Losses

```
# Grouping data by Season and Team Name for Home Team Data and Away Team Data
# and then using the UNION function to combine both the datasets
seasonwise_pts_df <- union(
    df %>%
    group_by(Season, HomeTeam) %>%
    dplyr::summarize(
        Pts = sum(HT_Pts_Earned),
        Goals_Scored = sum(FTHG),
        Goals_Conceeded = sum(FTAG),
        Wins = sum(HT_Win),
        Draws = sum(HT_Draw),
        Losses = sum(HT_Loss)) %>%
    as.data.frame() %>%
      rename(TeamName = HomeTeam),
    df %>%
    group_by(Season, AwayTeam) %>%
    dplyr::summarize(
        Pts = sum(AT_Pts_Earned),
        Goals_Scored = sum(FTAG),
        Goals_Conceeded = sum(FTHG),
        Wins = sum(AT_Win),
        Draws = sum(AT_Draw),
```

```
        Losses = sum(AT_Loss)) %>%
    as.data.frame() %>%
    rename(TeamName = AwayTeam))

# Grouping again by TeamName and Season to summate the values for the home and away games
seasonwise_pts_df <-  seasonwise_pts_df %>%
        group_by(Season, TeamName) %>%
          dplyr::summarize(Pts = sum(Pts),
          Goals_Scored = sum(Goals_Scored),
          Goals_Conceeded = sum(Goals_Conceeded),
          Wins = sum(Wins),
          Draws = sum(Draws),
          Losses = sum(Losses)
        ) %>%
        as.data.frame()
```

3. Creating derived columns such as Goal Difference(Goals Scored - Goals conceded) from the seasonwise results dataset

```
seasonwise_pts_df$Goal_Difference <-
  seasonwise_pts_df$Goals_Scored - seasonwise_pts_df$Goals_Conceeded
```

4. Obtaining the league position in each season using the rank function(Rank Priority - Points Earned> Goal Difference, Goals Scored)

```
seasonwise_pts_df <- seasonwise_pts_df %>%
    group_by(Season) %>%
    mutate(Rank = order(
      order(Pts,
            Goal_Difference,
            Goals_Scored,
            decreasing=TRUE)))
```

Problem 2

Analysing CHELSEA's performance in the English Premier League

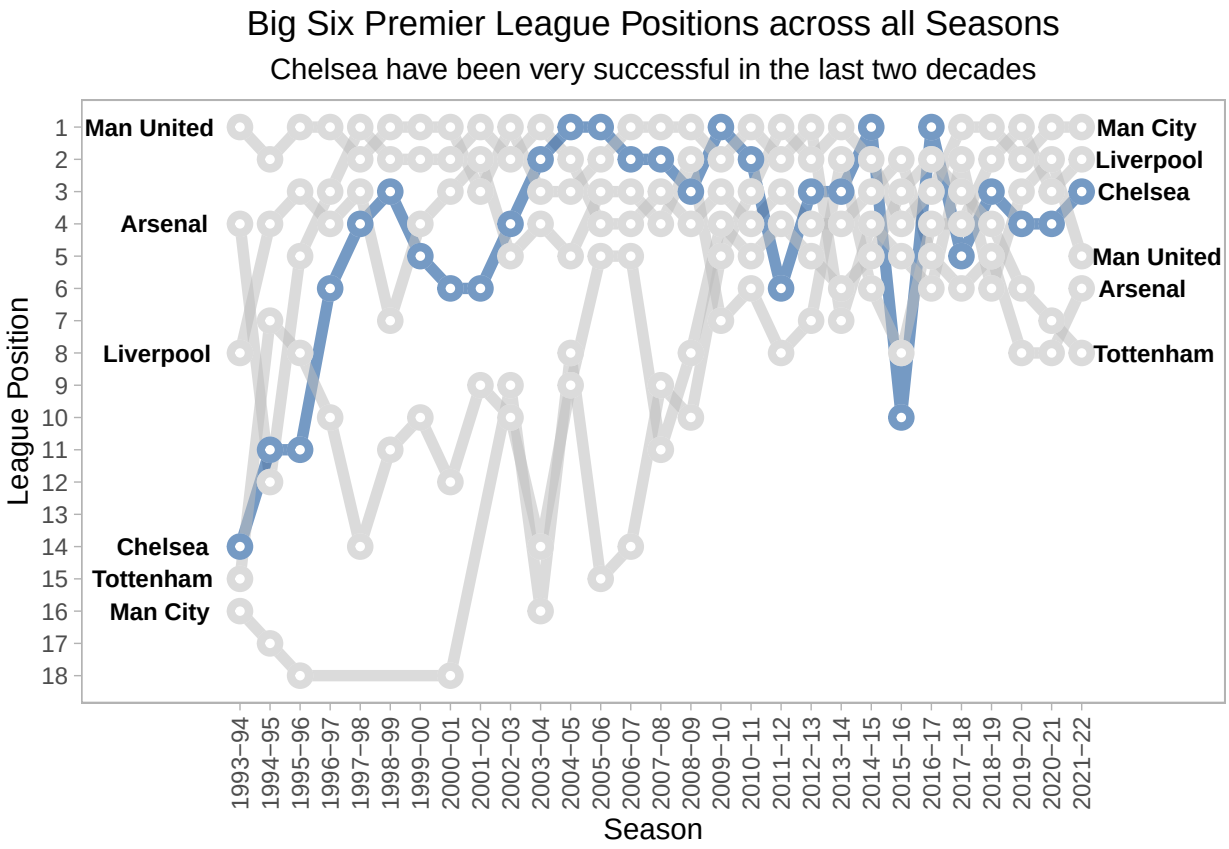
1. Chelsea's league performance in the Premier League

Instead of comparing Chelsea to every other team in the League, it would be much better to compare with just the consistently performing teams. The top six teams in the past decade also known as the **Big Six** have been Arsenal,Chelsea, Manchester City, Manchester United, Liverpool and Tottenham. So let us compare Chelsea's perfromance with the rest of the Big Six

```
#Subset the seasonwise results for the Big Six
bigsix <- filter(seasonwise_pts_df,
                 TeamName %in% c("Arsenal",
                                 "Chelsea",
```

```
                                "Man City",
                                "Man United",
                                "Liverpool",
                                "Tottenham"))

#Creating a flag to use to highlight Chelsea's performance
bigsix <- bigsix %>%
  mutate(flag = ifelse(TeamName %in% c("Chelsea"),
                       TRUE,
                       FALSE),
         club_col = if_else(flag == TRUE,
                            TeamName,
                            "zzz"))

#Creating a bump chart to show rankings
ggplot(data = bigsix,
       aes(x = as.numeric(factor(bigsix$Season)) ,
       y = Rank,
       group = TeamName)) +
  geom_line(aes(color = club_col, alpha = 1), size = 2) +
  geom_point(color = "#FFFFFF", size = 4) +
  geom_point(aes(color = club_col, alpha = 1), size = 4) +
  geom_point(color = "#FFFFFF", size = 1) + theme_light() +
  scale_y_reverse(breaks = 1:nrow(bigsix)) +
  scale_x_continuous(breaks = 1:29,
                     labels = levels( factor(bigsix$Season)),
                     minor_breaks = 1:29,
                     expand = c(.12, .12)) +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust=.5)) +
  geom_text(data = bigsix %>% filter(Season == "1993-94"),
            aes(label = TeamName,
                x = - 0.5) ,
            hjust = .85,
            fontface = "bold",
            color = "#000000",
            size = 3) +
  geom_text(data = bigsix %>% filter(Season == "2021-22"),
            aes(label = TeamName,
                x = 30) ,
            hjust = 0.15,
            fontface = "bold",
            color = "#000000",
            size = 3) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = "none",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(x = "Season",
       y = "League Position",
       title = "Big Six Premier League Positions across all Seasons",
       subtitle = "Chelsea have been very successful in the last two decades")  +
```

```
  scale_color_manual(values = c("#034694","grey"))
```



Observations

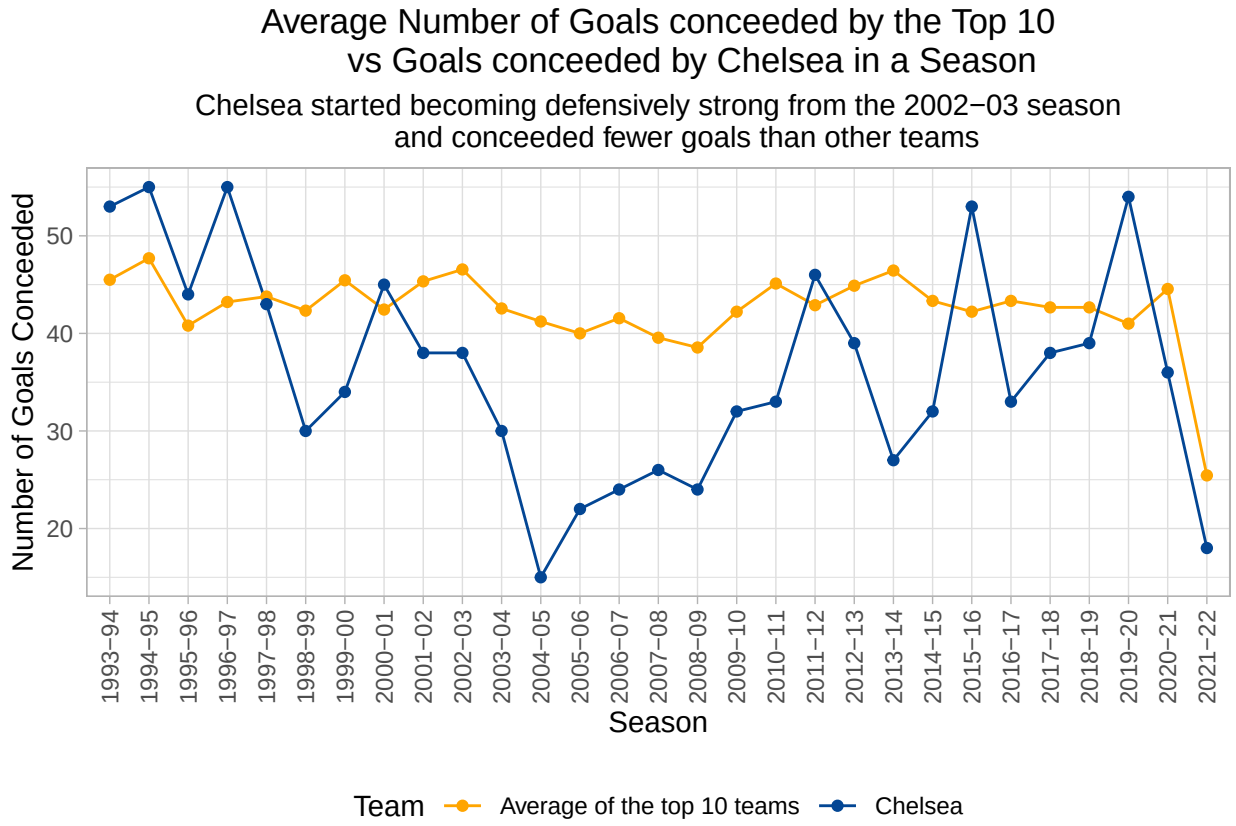
- From the graph, we see Chelsea's success at a high from the 2004-05 season
  - This quality of growth could be attributed to **Roman Abramovich** buying the club in **2003** and investing heavily into the team
- Other than an anomaly of the 2015-16 Season, we see an upward trajectory in Chelsea's performances within the past two decades
  - This **Ownership Impact** is as the media call it, *The Abramovich effect*

2. Analysing CHELSEA's statwise performance in the English Premier League

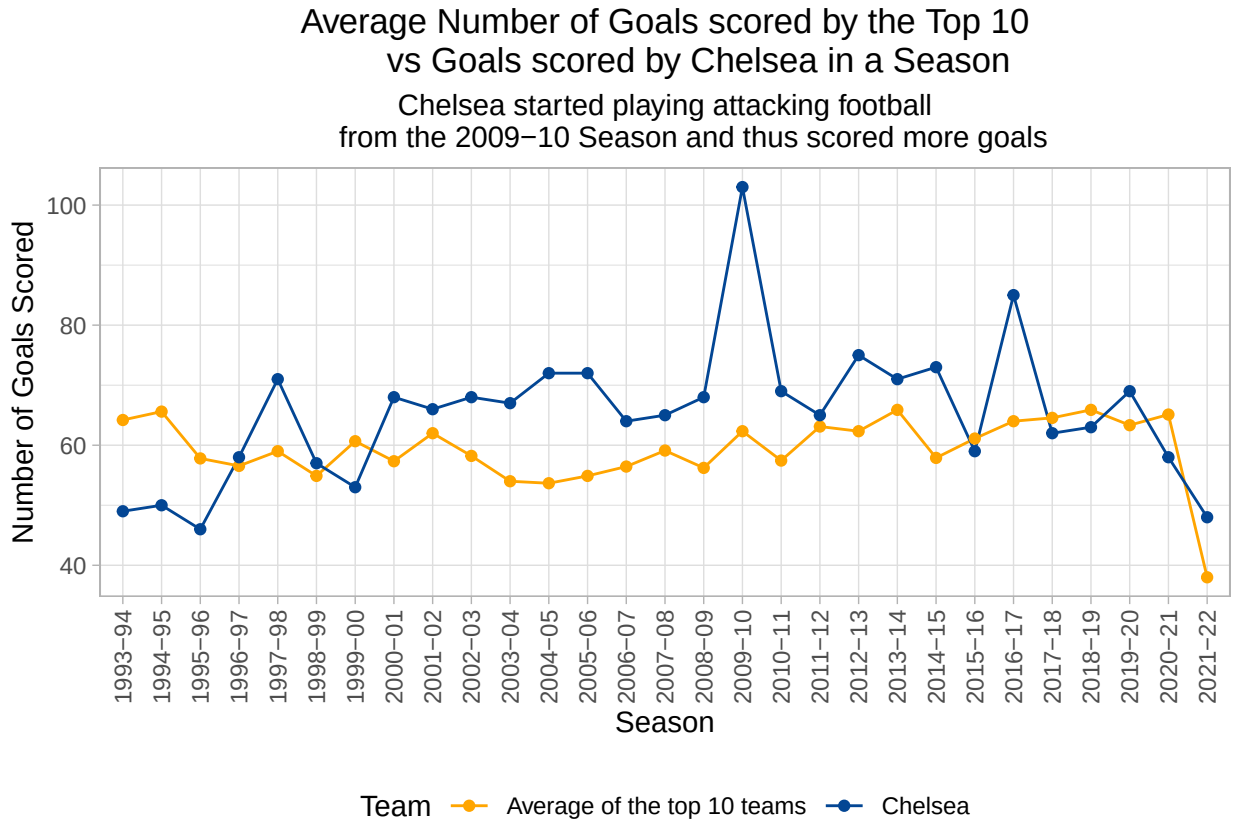
Comparing Chelsea's stats with the stats of the Top 10 teams of each Season (excluding Chelsea) to see how Chelsea have fared in attack and defence.

```
#Filtering the top 10 teams of each Season(excl Chelsea)
#based on Rank and calculating the mean of their stats

average_topten_performance <- filter(
  seasonwise_pts_df,
  TeamName != 'Chelsea' & Rank <= 10)%>%
```

```
    group_by(Season) %>%
    dplyr::summarize(Goals_Scored = mean(Goals_Scored),
                     Goals_Conceeded = mean(Goals_Conceeded),
                     Wins = mean(Wins),
                     Draws = mean(Draws),
                     Losses = mean(Losses)) %>%
    as.data.frame()

#Creating a new column to store Team Name
average_topten_performance$Team <- 'Average of the top 10 teams'

#Filtering out Chelsea's seasonwise performance
chelsea_performance <- filter(
  seasonwise_pts_df, TeamName == 'Chelsea') %>%
    group_by(Season) %>%
    dplyr::summarize(Goals_Scored = sum(Goals_Scored),
                     Goals_Conceeded = sum(Goals_Conceeded),
                     Wins = sum(Wins),
                     Draws = sum(Draws),
                     Losses = sum(Losses)) %>%
    as.data.frame()
#Creating a new column to store Team Name
chelsea_performance$Team <- 'Chelsea'

#Combining both the dataframes to include Chelsea's stats and the #average stats of the top 10 in the l
performance_wise_df <- union(average_topten_performance,chelsea_performance)

#Creating line charts to compare the defence and attack of the top 10 to Chelsea


#Plotting Goals Conceded per Season by Chelsea and the Average
#of the goals scored by the Top 10
ggplot(performance_wise_df,
       aes(x = Season,
           y = Goals_Conceeded,
           group=Team)) +
  geom_line(aes(color=Team))+
  geom_point(aes(color=Team)) +
  theme_light() +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust=.5)) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = "bottom") +
  labs(x = "Season",
       y = "Number of Goals Conceeded",
       title = "Average Number of Goals conceeded by the Top 10
       vs Goals conceeded by Chelsea in a Season",
       subtitle = "Chelsea started becoming defensively strong from the 2002-03 season
       and conceeded fewer goals than other teams")  + scale_color_manual(values =
                    c("Average of the top 10 teams" = "orange",
                      Chelsea = "#034694"))
```




```
#Plotting Goals Scored per Season by Chelsea and the Average of the goals scored by the Top 10
ggplot(performance_wise_df,
       aes(x = Season,
           y = Goals_Scored,
           group = Team)) +
  geom_line(aes(color = Team)) +
  geom_point(aes(color = Team)) + theme_light() +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.5)) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = "bottom") +
  labs(x = "Season",
       y = "Number of Goals Scored",
       title = "Average Number of Goals scored by the Top 10
       vs Goals scored by Chelsea in a Season",
       subtitle = "Chelsea started playing attacking football
       from the 2009-10 Season and thus scored more goals") + scale_color_manual(values =
                  c("Average of the top 10 teams" = "orange",
                    Chelsea = "#034694"))
```



Observations

- **Defensive Statistics** - From plot 1, we see a general trend where Chelsea seem to concede fewer goals than the average of the top 10 teams, in the past two decades
  - Chelsea conceded just 15 goals in the 2004-05 Season
  - This was due to the new defensive tactics employed by then Manager, Jose Mourinho
- **Attacking Statistics** - From plot 2, we see that Chelsea generally scores more than the average of the top 10 teams
  - We see a meteoric rise in goals scored by Chelsea from the 2008-09 Season to the 2009-10 season
  - Carlo Ancelotti, Chelsea's Manager at that time, implemented attacking tactics that led to this increase in goals scored
- These observations show the importance of Managerial Tactics