HOMEWORK 1

Deadline: 27/09/2022

Anna Martirosyan

Home advantage

Overall 100 pts (including bonus question).
Submit both markdown and pdf files

1. Calculate average number of goals for home and away teams per SEASON for that league, use dplyr.

```
new_data = f_data_sm  %>%
  filter(COUNTRY=='Turkey') %>%
  group_by(SEASON) %>%
  summarise(mean(FTHG), mean(FTAG))
new_data
```

```
## # A tibble: 28 x 3
##    SEASON `mean(FTHG)` `mean(FTAG)`
##     <dbl>        <dbl>        <dbl>
##  1   1995         1.90         1.18
##  2   1996         1.69         1.16
##  3   1997         1.69         1.17
##  4   1998         1.69         1.12
##  5   1999         1.68         1.18
##  6   2000         1.70         1.17
##  7   2001         1.92         1.40
##  8   2002         1.77         1.20
##  9   2003         1.63         1.16
## 10   2004         1.66         1.30
## # ... with 18 more rows
```

2. Calculate average number of goals for home and away teams per SEASON for that league for both top and bottom teams, use dplyr. Is there any difference in between the top and bottom teams in terms of average number of goals ? The function get_top_bottom_teams can be used to get the top and bottom teams (10p)

```
top_teams=get_top_bottom_teams(f_data_sm , "Turkey", top=TRUE)
bottom_teams = get_top_bottom_teams(f_data_sm , "Turkey", top=FALSE)

top_team_results= top_teams %>%
  group_by(SEASON) %>%
  summarise(TOP_MEAN_FTAG = mean(FTAG), TOP_MEAN_FTHG = mean(FTHG))

bottom_teams_results = bottom_teams %>%
```

```
  group_by(SEASON) %>%
  summarise(BOTTOM_MEAN_FTAG = mean(FTAG), BOTTOM_MEAN_FTHG= mean(FTHG))

final_df = inner_join(top_team_results, bottom_teams_results,
                              by = "SEASON")
final_df
```

```
## # A tibble: 28 x 5
##    SEASON TOP_MEAN_FTAG TOP_MEAN_FTHG BOTTOM_MEAN_FTAG BOTTOM_MEAN_FTHG
##     <dbl>         <dbl>         <dbl>            <dbl>            <dbl>
##  1   1995          1.19          1.81             1.04             1.87
##  2   1996          1.19          1.65             1.09             1.48
##  3   1997          1.08          1.71             1.02             1.63
##  4   1998          1.21          1.71            0.833             1.68
##  5   1999          1.32          1.32            0.911             1.97
##  6   2000          1.28          1.62             1.01             1.76
##  7   2001          1.18          2.03             1.52             1.89
##  8   2002          1.35          1.85              0.9             1.66
##  9   2003          1             1.64             1.17              1.6
## 10   2004          1.39          1.64             1.14             1.74
## # ... with 18 more rows
```

```
# From the result we can see a slightly difference in average number of goals
# for home and away teams classified as tops and bottoms. Top teams' average
# FTAG nearly always exceeds bottom teams' and ranges from 0.9-1.38, but some
# exceptions are also  possible, for example for year 2010. The same could be
# said for FTHG, mean of  top teams is mostly higher than mean of bottom teams,
# but for some years, for example 2004, mean of bottom teams exceeds the mean
# of top teams.
```

3. Now calculate the same statistics for the games where the home team was in bottom of the table and the away team was in the top of the table. Is the home advantage still a strong factor ? (10p)

```
n_data=get_top_vs_bottom_teams(f_data_sm, 'Turkey')
FTHG = get_top_vs_bottom_teams(f_data_sm, 'Turkey')$FTHG
FTAG = get_top_vs_bottom_teams(f_data_sm, 'Turkey')$FTAG
n_data  %>%
  group_by(SEASON) %>%
  summarise(mean(FTHG), mean(FTAG))
```

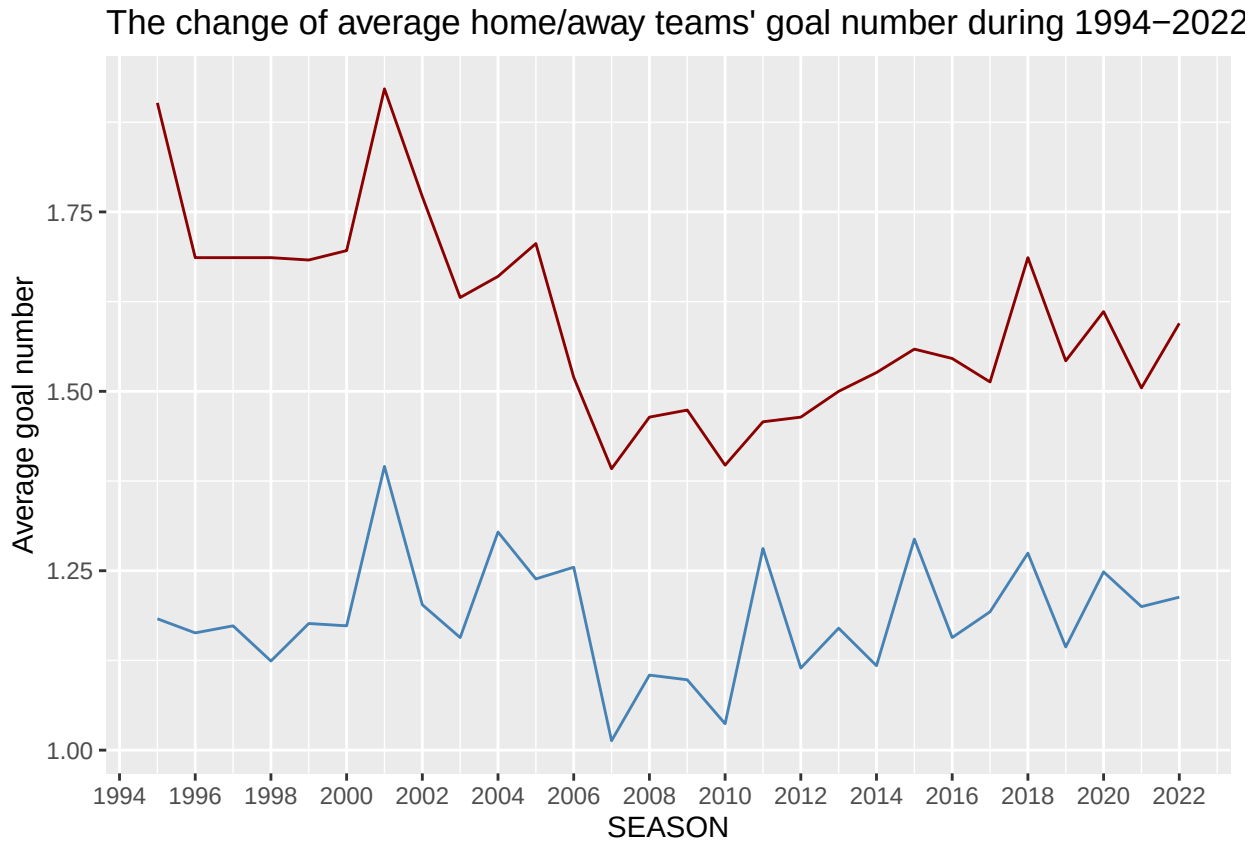
```
## # A tibble: 28 x 3
##    SEASON `mean(FTHG)` `mean(FTAG)`
##     <dbl>        <dbl>        <dbl>
##  1   1995         1.27         1.79
##  2   1996         1.21         1.61
##  3   1997        0.910         1.91
##  4   1998         1.21         1.46
##  5   1999         1.13         1.62
##  6   2000         1.25         1.71
##  7   2001         1.44         1.79
##  8   2002         1.43         1.55
##  9   2003         1.11         1.73
```

```
## 10   2004        1.18          1.44
## # ... with 18 more rows
```

```
# From the below data frame could be seen that mean of home goals is always
# less than the mean of away goals so the home advantage is not still a strong
# factor.
```

4. Construct a plot using ggplot to show how this number is changing over time. Use the means from the first exercise. Note: you need to have SEASON on x-axis. Show average Home goals and Away goals on the same plot. Be sure that your plot has appropriate axis names and title (10p)

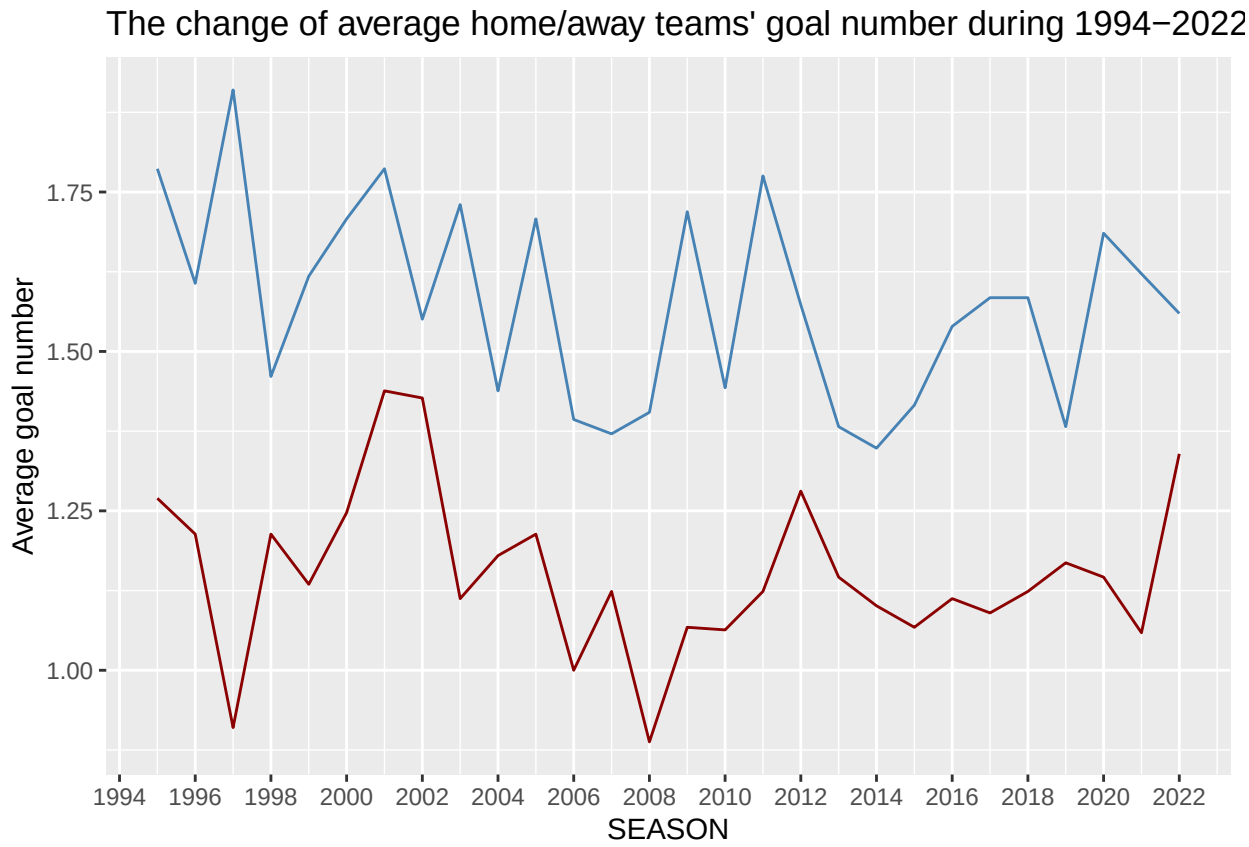
```
library(ggplot2)
ggplot(new_data, aes(x=SEASON)) +
  geom_line(aes(y = `mean(FTHG)`), color = "darkred") +
  geom_line(aes(y = `mean(FTAG)`), color="steelblue") +
  ylab("Average goal number")+
  scale_x_continuous(breaks = seq(1994, 2022, 2))+
  ggtitle("The change of average home/away teams' goal number during 1994-2022")
```



5. Now use the same plot on the games where the away team was from top of the table (10p)

```
new_new_data = n_data  %>%
  group_by(SEASON) %>%
  summarise(mean(FTHG), mean(FTAG))
```

```
ggplot(new_new_data, aes(x=SEASON)) +
  geom_line(aes(y = `mean(FTHG)`), color = "darkred") +
  geom_line(aes(y = `mean(FTAG)`), color="steelblue") +
  ylab("Average goal number")+
  scale_x_continuous(breaks = seq(1994, 2022, 2))+
  ggtitle("The change of average home/away teams' goal number during 1994-2022")
```

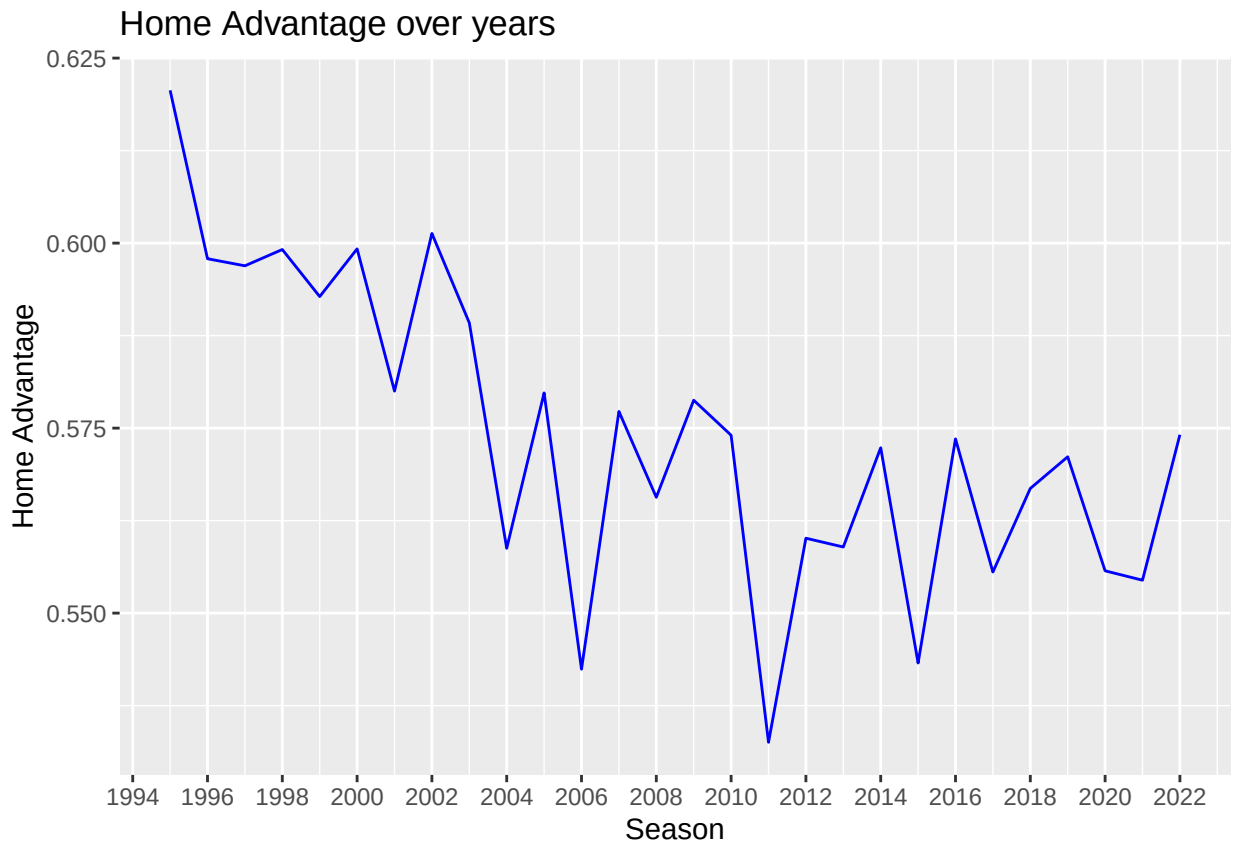


6. Interpret the plots (10p)

```
# From the 2 plots below we can see that  on the games where the away team was
# from top of the table the average number of home goals decreased, initially
# it was in the range 1.3-1.9 but if away team was from top the range changed
# from 0.8-1.4. This results are quite expected, because if home team is from
# bottom of the table, it means that it is not a "very good" team, so the
# average of home points should be lower.
#
# The range of away points in the first case was 1.01-1.39, but if away team is
# in the top the range changes from 1.3-1.9. Which is also expected, because
# if away team is from top teams (is one of the best teams), that its average
# points should be higher compared to the points in problem 4.
```

7. Think of your own approach on how will you measure home team advantage given the data you have. Calculate Home team advantage for your league and for all other leagues over time (10p)

```r
######################  Home advantage for my league ######################
won_count_home_team <- f_data_sm %>%
  filter(COUNTRY=='Turkey')%>%
  group_by(HOMETEAM, SEASON) %>%
  summarize(TOTAL_HP = sum(FTHG)) %>%
  rename(TEAM_NAME = HOMETEAM)

won_count_away_team <- f_data_sm %>%
  filter(COUNTRY=='Turkey')%>%
  group_by(AWAYTEAM, SEASON) %>%
  summarize(TOTAL_AP = sum(FTAG)) %>%
  rename(TEAM_NAME = AWAYTEAM)

statistics <- inner_join(won_count_home_team, won_count_away_team,
                             by = c("TEAM_NAME", "SEASON")) %>% distinct()

statistics_with_portion_1 <- statistics %>%
  mutate(portion_1 = TOTAL_HP / (TOTAL_HP + TOTAL_AP)) %>%
  rename_with(toupper)

statistics_by_season_1 <- statistics_with_portion_1 %>%
  group_by(SEASON) %>%
  summarise(PORTION_1 = mean(PORTION_1))

ggplot(statistics_by_season_1,aes(x = SEASON, group = 1)) +
geom_line(aes(y = PORTION_1), color = "blue") +
  ggtitle("Home Advantage over years") +
  xlab('Season') + ylab('Home Advantage')+
  scale_x_continuous(breaks = seq(1994, 2022, 2))
```




```
# Below you can see a graph which a clear interpretation of home advantage.
# By taking a closer look at the numbers we can see, that for example in
# 2002  approximately 72% of total points are made by home team. The result is
# nearly the same for other years as well. From the graph it can be seen that
# the lowest point is at 2011, but even in that year approximately 68% of total
 #points are made by home team.


########################  Home advantage for all leagues ########################

won_count_home_team <- f_data_sm %>%
  group_by(HOMETEAM, SEASON, COUNTRY) %>%
  summarize(TOTAL_HP = sum(FTHG)) %>%
  rename(TEAM_NAME = HOMETEAM)

won_count_away_team <- f_data_sm %>%
  group_by(AWAYTEAM, SEASON, COUNTRY) %>%
  summarize(TOTAL_AP = sum(FTAG)) %>%
  rename(TEAM_NAME = AWAYTEAM)

statistics <- inner_join(won_count_home_team, won_count_away_team,
                         by = c("TEAM_NAME", "SEASON", "COUNTRY")) %>%
  distinct()

statistics_with_portion_2 <- statistics %>%
  mutate(HOME_ADVANTAGE = TOTAL_HP / (TOTAL_HP + TOTAL_AP)) %>%
```

```
  rename_with(toupper)

statistics_by_season_2 <- statistics_with_portion_2 %>%
  group_by(SEASON, COUNTRY) %>%
  summarise(HOME_ADVANTAGE = mean(HOME_ADVANTAGE))

statistics_by_season_2
```

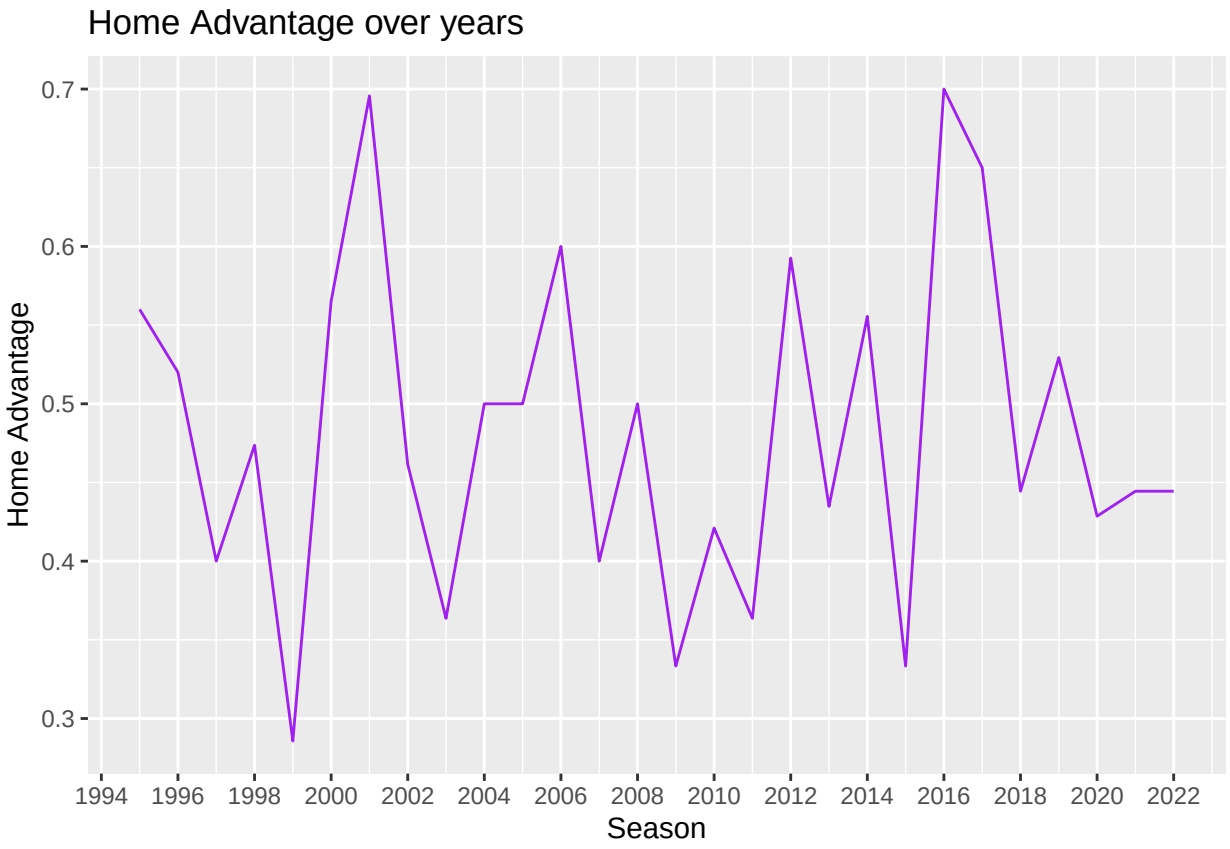
```
## # A tibble: 313 x 3
## # Groups:   SEASON [29]
##    SEASON COUNTRY     HOME_ADVANTAGE
##     <dbl> <chr>                <dbl>
##  1   1994 England              0.558
##  2   1994 France               0.634
##  3   1994 Germany              0.600
##  4   1994 Italy                0.610
##  5   1994 Netherlands          0.588
##  6   1994 Spain                0.606
##  7   1995 England              0.585
##  8   1995 France               0.645
##  9   1995 Germany              0.590
## 10   1995 Greece               0.665
## # ... with 303 more rows
```

```
# By opening the data frame statistics_by_season you can see a column named
# HOME_ADVANTAGE, which is an indicator of home advantage, if the number in
# HOME_ADVANTAGE is closer to 1, it means that home advantage is quite big. And
# if HOME_ADVANTAGE is closer to 0, it means that home advantage is
# unnoticeable.
```

8. Calculate the same ratio for the games where the away team was from the top of the table and the home team was from the bottom of the table. (10p)

```
won_count_home_team <- n_data %>%
  group_by(HOMETEAM, SEASON) %>%
  summarize(TOTAL_HP = sum(FTHG)) %>%
  rename(TEAM_NAME = HOMETEAM)

won_count_away_team <- n_data %>%
  group_by(AWAYTEAM, SEASON) %>%
  summarize(TOTAL_AP = sum(FTAG)) %>%
  rename(TEAM_NAME = AWAYTEAM)

statistics <- inner_join(won_count_home_team, won_count_away_team,
                         by = c("TEAM_NAME", "SEASON")) %>% distinct()


statistics_with_portion_3 <- statistics %>%
  mutate(portion_3 = TOTAL_HP / (TOTAL_HP + TOTAL_AP)) %>%
  rename_with(toupper)

statistics_by_season_3 <- statistics_with_portion_3 %>%
```

```
  group_by(SEASON) %>%
  summarise(PORTION_3 = mean(PORTION_3))

ggplot(statistics_by_season_3, aes(x = SEASON, group = 1)) +
geom_line(aes(y = PORTION_3),  color = "purple") +
  ggtitle("Home Advantage over years") +
  xlab('Season') + ylab('Home Advantage')+
  scale_x_continuous(breaks = seq(1994, 2022, 2))
```

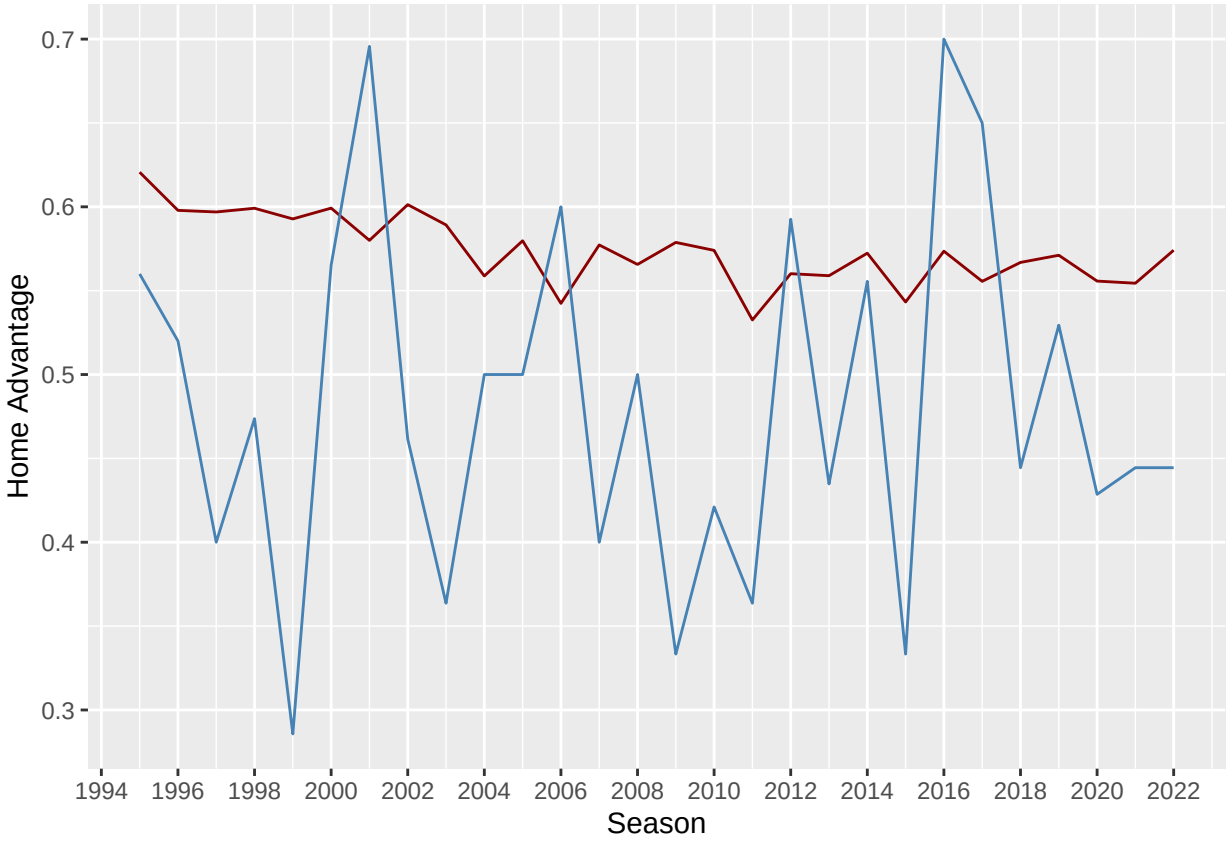



```
################################## Explanation ###################################

# The results of below graph are quite expected. Since the home team was from
# the bottom of the table it means that it's not a very strong team, which
# means that playing in their own field will increase their chances to win. So
# home advantage plays a huge factor for bottom teams. Whereas, for top teams,
# which are much more stronger teams than bottom ones, playing in their own
# field does not play a huge role. Below you can see the results graphically:
```

9. Plot the results for all games for your league and the games where the away team was from the top of the table in one plot using ggplot2. (10p)

```
overall = inner_join(statistics_by_season_1, statistics_by_season_3,
                                by = "SEASON")
ggplot(overall, aes(x=SEASON))+
```

```
  geom_line(aes(y = PORTION_1), color = "darkred") +
  geom_line(aes(y = PORTION_3), color = "steelblue") +
  xlab('Season') + ylab('Home Advantage')+
  scale_x_continuous(breaks = seq(1994, 2022, 2))
```



Predictions

1. With your chosen league, pick an upcoming game and calculate the betting odds for that game. Use distribution approach. You can compare the result with the actual betting odds from this website, https://www.bet365.com/#/AS/B1/. (10p)

```
home = f_data_sm %>%
                filter(COUNTRY=="Turkey") %>%
                group_by(HOMETEAM) %>%
                summarise(mean = mean(FTHG))
away = f_data_sm %>%
                filter(COUNTRY=="Turkey") %>%
                group_by(AWAYTEAM) %>%
                summarise(mean = mean(FTAG))

erzurum = home$mean[home$HOMETEAM=="Erzurum BB"]
denizlispor = away$mean[away$AWAYTEAM=="Denizlispor"]

options(scipen=999)
```

```r
goal_probs_erzurum=dpois(c(0:9), lambda=erzurum)
goal_probs_denizlispor=dpois(c(0:9), lambda=denizlispor)

options(width = 300)
matrix_data=goal_probs_erzurum %*% t(goal_probs_denizlispor)
matrix_data=round(matrix_data, digits=4)

#Probability of Antalia to win
erzurum_win = sum(matrix_data[lower.tri(matrix_data, diag=F)])

# Probability of Adanaspor to win
denizlispor_win =  sum(matrix_data[upper.tri(matrix_data, diag=F)])

# Probability of draw
draw = sum(diag(matrix_data))

(odds_for_erzurum_win = 1/erzurum_win)
```

```
## [1] 3.081664
```

```r
(odds_for_denizlispor_win = 1/denizlispor_win)
```

```
## [1] 2.698327
```

```r
(odds_for_draw = 1/draw)
```

```
## [1] 3.279764
```

Power Ratings

1. Try predicting home results with power ratings in other sports, Does it work as well as in football ? Try to interpret the actual results. The data for nba, nfl is available in SportsAnalytics270 (Bonus question, 10 points)