Data Exercise CH3 Q3

Fasih Atif

10/23/2020

Install required libraries

```
library(tidyverse)
library(lattice)
library(Hmisc)
```

Import the football dataset

```
data_in <- "C:/Users/Atif_Fasih/Downloads/"
epl_games <- read_csv(paste0(data_in,"epl_games.csv"))
```

Filter dataset for 2018 season

```
epl_2018 <- epl_games %>% filter(season == 2018)
```

Create a goal_diff variable

```
epl_2018 <- mutate(epl_2018, "goal_diff" = goals_home - goals_away)
```

Summary statistics for the goal_diff variable

```
Hmisc::describe(epl_2018$goal_diff)
```

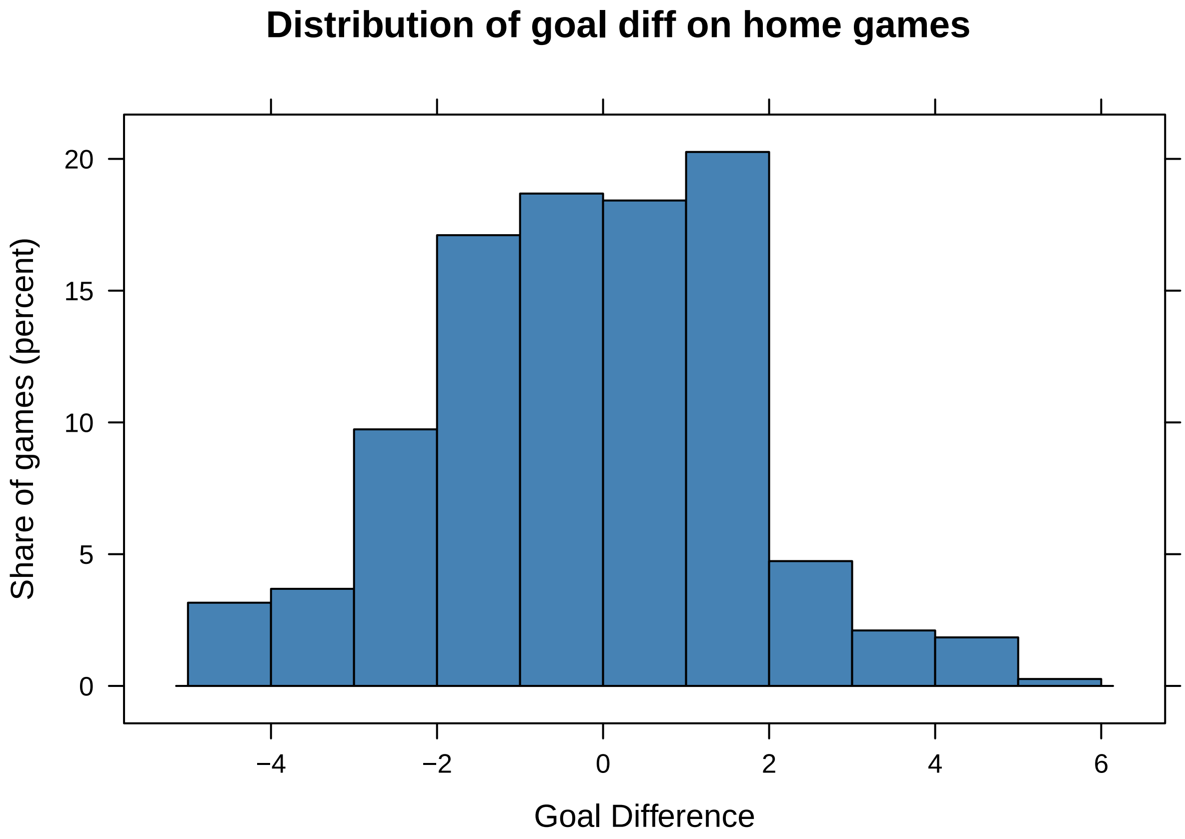
```
## epl_2018$goal_diff
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      380        0       12    0.973   0.3158    2.131       -3       -2
##      .25      .50      .75      .90      .95
##       -1        0        2        2        3
##
## lowest : -5 -4 -3 -2 -1, highest:  2  3  4  5  6
##
## Value         -5    -4    -3    -2    -1     0     1     2     3     4     5
## Frequency      2    10    14    37    65    71    70    77    18     8     7
## Proportion 0.005 0.026 0.037 0.097 0.171 0.187 0.184 0.203 0.047 0.021 0.018
##
## Value          6
## Frequency      1
## Proportion 0.003
```

```
summarise(epl_2018, "std" = sd(epl_2018$goal_diff))
```

```
## # A tibble: 1 x 1
##     std
##   <dbl>
## 1  1.92
```

Create relative frequency histogram

```
histogram(epl_2018$goal_diff,
          main = "Distribution of goal diff on home games",
          xlab = "Goal Difference",
          ylab = "Share of games (percent)",
          col = "steelblue",
          breaks = 8)
```



Analysis

We are going to analyze the 2018-2019 football season.

We had two quantitative variables ‘goals_home’ and ‘goals_away’ from which we created another “goal_diff” variable in order to analyze whether teams have an increased scoring proabibility on home turf and hence whether they win the match.Since the goal difference doesnt have too many values, we show a histogram that shows the percentage of each value instead of bins.

The mode is goal difference of 2 with 20.3%. This means that in 20.3% of the games, home team won with a positive 2 goal difference. The histogram is symmetrical on both sides of zero to an extent.Since the proportion of positive goal differences 1 and 2 is more than -1 and -2, this suggests that the home team has a small advantage over the away team. On any randomly chosen 2018/2019 game, the home team is expected to score 0.31 goals more as shown by the mean.The Standard deviation is 1.9 which shows that the mean goal difference of 0.31 is neither neglible nor huge.

Overall, the home team had a winning percentage of 50.3% and the away team had a winning percentage of 31.6%.

Results of 2016-2017 season as taken from book:

Mean: 0.4 Std = 1.9 Percent Positive = 49 Percent Zero = 22 Percent Negative = 29

Comparing the results to the 2016/17 season, the teams in 2018 season had a better winning probability at home. Away teams also had a better winning chance (~3% more) in 2018 season compared to the 2016. The number of games drawn were 3.3% higher in the 2016/17 season.The mean remains nearly the same with 2018 season a tenth lower and standard deviation remains the same at 1.9.

To conclude, the overall trend remains the same in both seasons with the home team winning more matches than the away team.