Data Tasks

Chapter 2 - Data Exercise Q.5

Table showing the number of seasons each team spent in the English Premier League

```
## Parsed with column specification:
## cols(
##   div = col_character(),
##   season = col_double(),
##   date = col_character(),
##   team_home = col_character(),
##   team_away = col_character(),
##   points_home = col_double(),
##   points_away = col_double(),
##   goals_home = col_double(),
##   goals_away = col_double()
## )
```

Teams that played all 11 seasons in the EPL

```
seasons_played_in_epl %>% filter(No_Of_Seasons_In_Epl == 11) %>% select(Team)
```

```
## # A tibble: 7 x 1
## # Groups:   Team [7]
##   Team
##   <chr>
## 1 Arsenal
## 2 Everton
## 3 Chelsea
## 4 Man United
## 5 Liverpool
## 6 Tottenham
## 7 Man City
```

Teams that played only once in the EPL

```
seasons_played_in_epl %>% filter(No_Of_Seasons_In_Epl == 1) %>% select(Team)
```

```
## # A tibble: 2 x 1
## # Groups:   Team [2]
##   Team
##   <chr>
## 1 Blackpool
## 2 Reading
```

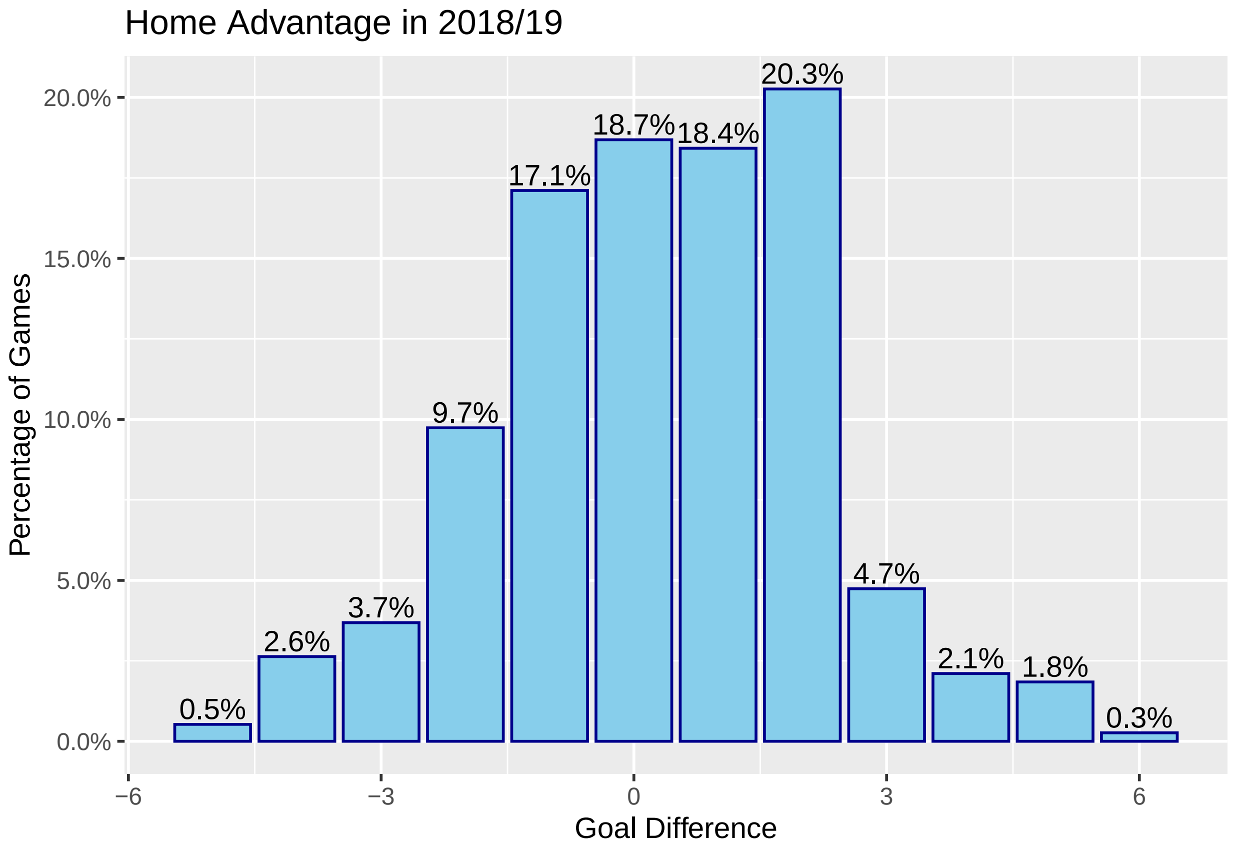
Chapter 3 - Data Exercise Question 3

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
home_adv_2018 <- epl_data %>% filter(season == 2018)


  home_adv_2018 %>% ggplot(mapping = aes(x = goals_home - goals_away)) + geom_bar(aes(y = (..count..)/su
  geom_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))),
  scale_y_continuous(labels = percent) + labs(title = "Home Advantage in 2018/19", y = "Percentage of Ga
```



```
home_adv_2018 %>% summarise(No_of_observations = n(),
                              Mean = round(mean(goals_home - goals_away), 2),
                              standard_dev = round(sd(goals_home - goals_away), 2),
                              Percent_positive = round(sum(goals_home - goals_away > 0)/n(), 2)*100,
                              Percent_zero = round(sum(goals_home - goals_away == 0)/n(), 2)*100,
                              Percent_negative = round(sum(goals_home - goals_away < 0)/n(), 2)*100)
```

```
## # A tibble: 1 x 6
##   No_of_observati~  Mean standard_dev Percent_positive Percent_zero
##              <int> <dbl>        <dbl>            <dbl>        <dbl>
## 1              380  0.32         1.92               48           19
## # ... with 1 more variable: Percent_negative <dbl>
```

Conclusion: We plotted a frequency distribution of Goal Difference in the 2018/19 season and found out that the mode was higher in 2018 compared to 2017 (in the book). The mean and standard deviation are also slightly higher in 2018, while the percentage of matches with negative goal difference is also higher in 2018 and the percentage of matches with zero goal difference is lower in 2018 compared to 2016.

Chapter 4 - Data Exercise Question 2.

```
home_totals <- epl_data %>%
  filter(season == 2017) %>%
  group_by(team_home) %>%
  summarise(total_points = sum(points_home)) %>%
  arrange(total_points) %>%
  rename(Team = team_home)

away_totals <- epl_data %>%
  filter(season == 2017) %>%
  group_by(team_away) %>%
  summarise(total_points = sum(points_away)) %>%
  arrange(total_points) %>%
  rename(Team = team_away)

season_totals <- merge(home_totals, away_totals, by = "Team")
season_totals <- season_totals %>%
  mutate(total_points = total_points.x + total_points.y) %>%
  arrange(desc(total_points))

season_2017_binned <- mutate(season_totals,
                             bin = cut(season_totals$total_points, c(-Inf, 54, 42, Inf),
                                       labels = c("Relegation Battle", "Mid-Table", "Top Six")))
season_2017_binned <- select(season_2017_binned, Team, bin)
season_2018 <- epl_data %>%
  filter(season == 2018) %>% rename(Team = team_home)
season_2018_binned <- merge(season_2017_binned, season_2018,
                            by = "Team", all.y = TRUE)
season_2018_binned$bin <- replace(season_2018_binned$bin,
                                  is.na(season_2018_binned$bin),
                                  "Relegation Battle")
```

2018 Statistics with Team sorted into three bins; Top Six, Mid-Table and Relegation Battle according to their in the previous season

```
season_2018_binned %>% distinct(Team, bin) %>% arrange(bin)
```

```
##              Team               bin
## 1        Brighton Relegation Battle
## 2         Cardiff Relegation Battle
## 3          Fulham Relegation Battle
## 4    Huddersfield Relegation Battle
## 5     Southampton Relegation Battle
## 6         Watford Relegation Battle
## 7        West Ham Relegation Battle
## 8          Wolves Relegation Battle
## 9     Bournemouth         Mid-Table
## 10        Burnley         Mid-Table
## 11 Crystal Palace         Mid-Table
## 12        Everton         Mid-Table
## 13      Leicester         Mid-Table
## 14      Newcastle         Mid-Table
## 15        Arsenal           Top Six
## 16        Chelsea           Top Six
## 17      Liverpool           Top Six
## 18       Man City           Top Six
## 19     Man United           Top Six
## 20      Tottenham           Top Six
```

```
season_2018_binned %>%
  group_by(bin) %>% summarise(No_of_observations = n(),
                              Mean = round(mean(goals_home - goals_away), 2),
                              standard_dev = round(sd(goals_home - goals_away), 2),
                              Percent_positive = round(sum(goals_home - goals_away > 0)/n(), 2)*100,
                              Percent_zero = round(sum(goals_home - goals_away == 0)/n(), 2)*100,
                              Percent_negative = round(sum(goals_home - goals_away < 0)/n(), 2)*100)
```

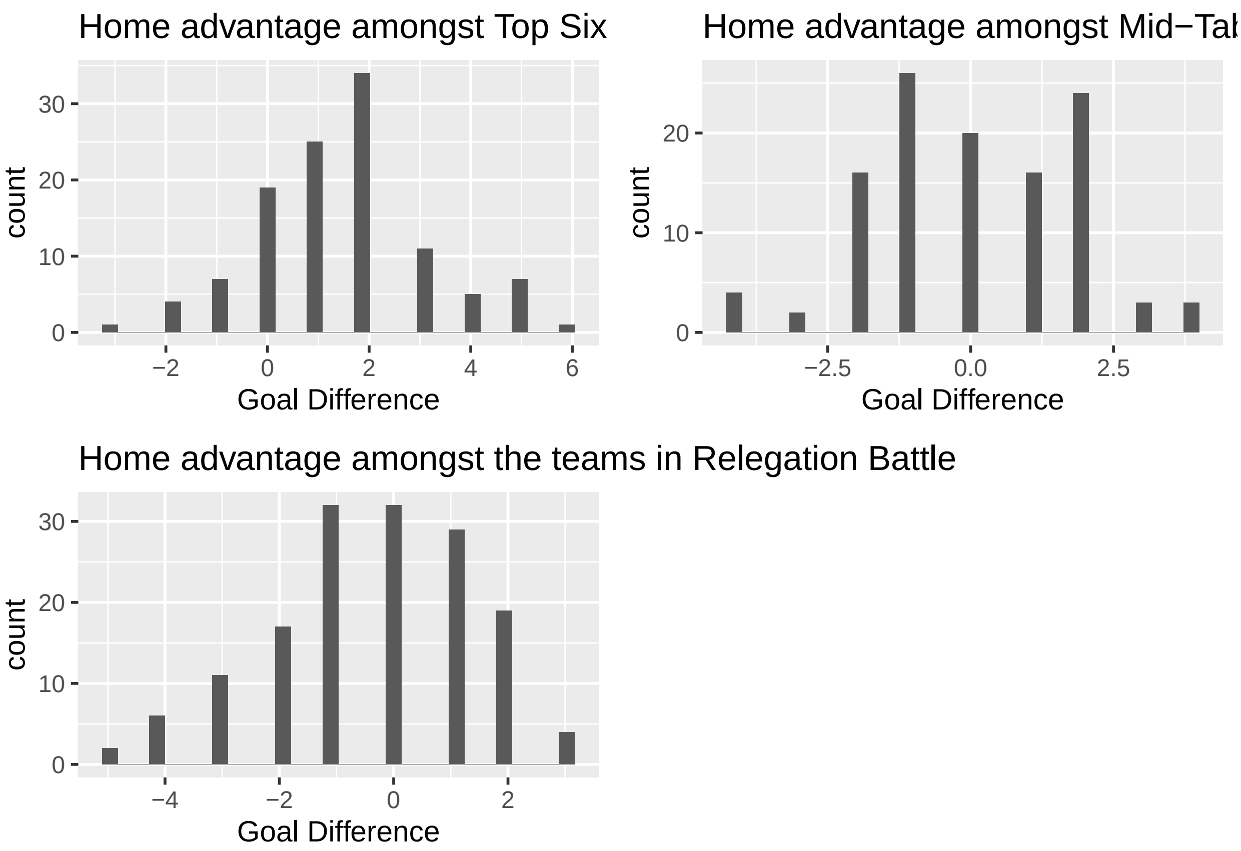
```
## # A tibble: 3 x 7
##   bin   No_of_observati~  Mean standard_dev Percent_positive Percent_zero
##   <fct>            <int> <dbl>        <dbl>            <dbl>        <dbl>
## 1 Rele~              152 -0.36         1.76               34           21
## 2 Mid--              114  0.04         1.78               40           18
## 3 Top ~              114  1.48         1.72               73           17
## # ... with 1 more variable: Percent_negative <dbl>
```

Home Advantage in our three bins

1. Histogram

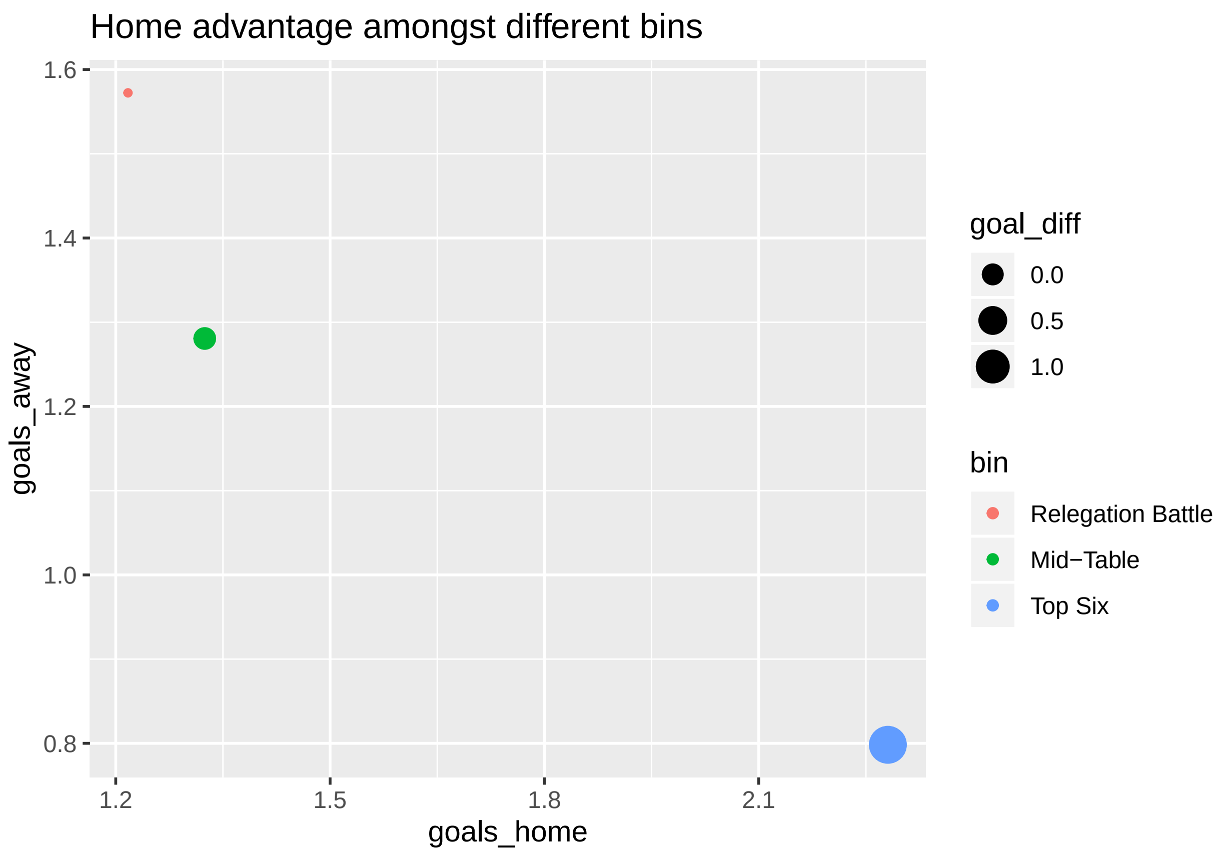
```
topSix <- season_2018_binned %>% filter(bin == "Top Six")

topSix_adv <- topSix %>%
  ggplot(mapping = aes(x = goals_home - goals_away)) +
  geom_histogram() +
  labs(title = "Home advantage amongst Top Six",
                        x = "Goal Difference")

midTable <- season_2018_binned %>% filter(bin == "Mid-Table")


midTable_adv <- midTable %>%
  ggplot(mapping = aes(x = goals_home - goals_away)) +
  geom_histogram() +
  labs(title = "Home advantage amongst Mid-Table",
                        x = "Goal Difference")

releBattle <- season_2018_binned %>% filter(bin == "Relegation Battle")

releBattle_adv <- releBattle %>%
  ggplot(mapping = aes(x = goals_home - goals_away))  +
  geom_histogram() +
  labs(title = "Home advantage amongst the teams in Relegation Battle",
                        x = "Goal Difference")

ggarrange(topSix_adv, midTable_adv, releBattle_adv, vjust = -3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2. Bin Scatter

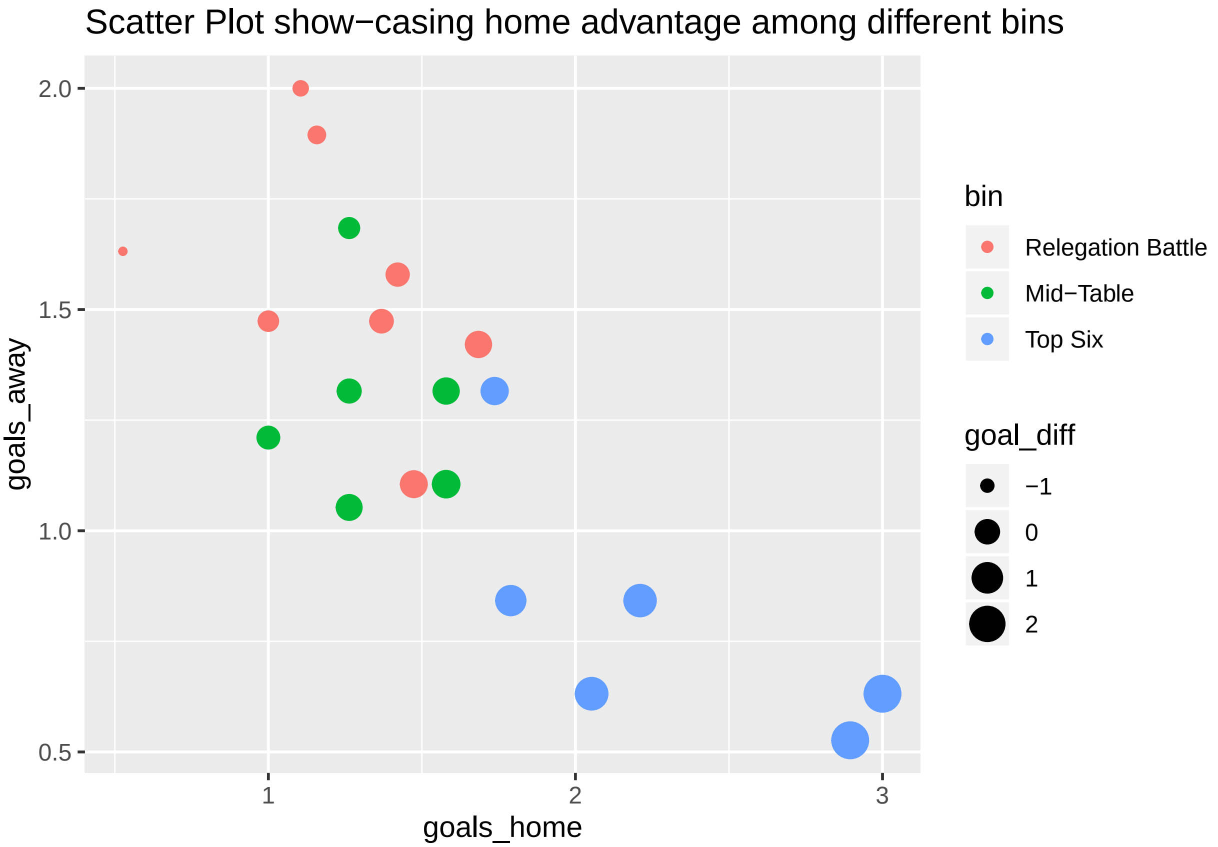
```
x <- season_2018_binned %>%
  group_by(bin) %>%
  summarise(goal_diff = mean(goals_home - goals_away),
            count_of_teams = n_distinct(Team), goals_home = mean(goals_home),
            goals_away = mean(goals_away))

x %>% ggplot(mapping = aes(x = goals_home , y = goals_away, color = bin, size = goal_diff)) +
  geom_point() + labs(title = "Home advantage amongst different bins")
```



3. Scatter Plot

```
y <- season_2018_binned %>%
  group_by(bin, Team) %>%
  summarise(goal_diff = mean(goals_home - goals_away),
            goals_home = mean(goals_home), goals_away = mean(goals_away))
y %>% ggplot(mapping = aes(x = goals_home , y = goals_away,
                           color = bin, size = goal_diff)) +
  geom_point(labels = season_2018_binned$Team) +
  labs(title = "Scatter Plot show-casing home advantage among different bins")
```

```
## Warning: Ignoring unknown parameters: labels
```



Conclusion: As can be seen by the two scatter plots, home advantage is most significant for the Top Six Teams, where they score around 2 home goals for each goal the away team scores. This advantage is very mild for mid_table teams where they score 1.32 home goals for every 1.28 away goals. Finally this pattern/advantage is inversed among teams that battle for relegation - these teams score 1 goal for about 1.5 every goal they concede at home.