Exercise Sheet 2

The *European Soccer Database* contains data on more than 25.000 national football matches from the best European leagues. The aim of this exercise is to present interesting relationships in R using explorative data analysis and visualization.

First you need to access some tables in the database. Note: You can use the `RSQLite::dbConnect()` function to do this. To access a particular database table and convert it to a `data.frame` , you can use the `tbl_df(dbGetQuery(connection, 'SELECT * FROM table_xyz'))` command.

```
# connect to database
con <- dbConnect(SQLite(), dbname = str_c(dirname(getwd()), "/VA/Data/EuropeanSoccer.sqlite"))
match <- tbl_df(dbGetQuery(con,"SELECT * FROM Match"))
league <- tbl_df(dbGetQuery(con,"SELECT * FROM League"))
```

1. The first leagues of Spain, England, Germany and Italy are considered the four most attractive football leagues in Europe.
    a. In which of the four leagues do on average score the most or the fewest goals per game?
    b. Compare the average, median, standard deviation, variance, range and interquartile distance of goals scored per match between the four most attractive European leagues and the remaining leagues.

```
match_top4 <- league %>%
  filter(name %in% c("Spain LIGA BBVA",
                     "England Premier League",
                     "Germany 1. Bundesliga",
                     "Italy Serie A")) %>%
  select(league_id = id, league_name = name) %>%
  inner_join(match, by = "league_id")

match_top4 %>%
  group_by(league_name) %>%
  filter(!is.na(home_team_goal) | !is.na(away_team_goal)) %>%
  summarize(avg_match_goals = mean(home_team_goal + away_team_goal)) %>%
  arrange(-avg_match_goals)
```

```
## # A tibble: 4 x 2
##   league_name            avg_match_goals
##   <chr>                            <dbl>
## 1 Germany 1. Bundesliga             2.90
## 2 Spain LIGA BBVA                   2.77
## 3 England Premier League            2.71
## 4 Italy Serie A                     2.62
```

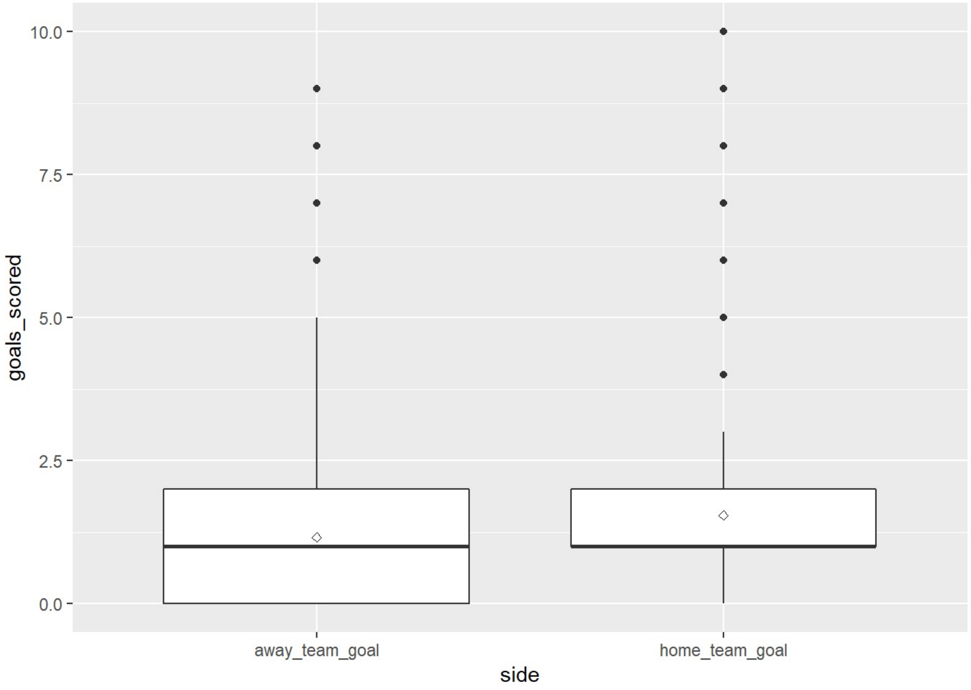
```
league %>%
  select(league_id = id, league_name = name) %>%
  inner_join(match, by = "league_id") %>%
  group_by(league_name) %>%
  filter(!is.na(home_team_goal) | !is.na(away_team_goal)) %>%
  summarize(avg_match_goals = mean(home_team_goal + away_team_goal)) %>%
  arrange(-avg_match_goals) %>%
  slice(c(1:5, (n()-4):n()))
```

```
league %>%
  mutate(name = fct_collapse(name,
                             top4 = c("Spain LIGA BBVA",
                                      "England Premier League",
                                      "Germany 1. Bundesliga",
                                      "Italy Serie A"))) %>%
  mutate(name = fct_other(name, keep = "top4", other_level = "rest")) %>%
  select(league_id = id, league_name = name) %>%
  inner_join(match, by = "league_id") %>%
  group_by(league_name) %>%
  mutate(match_goals = home_team_goal + away_team_goal) %>%
  summarise(
    avg_match_goals = mean(match_goals),
    median_match_goals = median(match_goals),
    sd_match_goals = sd(match_goals),
    var_match_goals = var(match_goals),
    min_match_goals = min(match_goals),
    max_match_goals = max(match_goals),
    iqr_match_goals = IQR(match_goals)
  ) %>% knitr::kable()
```

league_name	avg_match_goals	median_match_goals	sd_match_goals	var_match_goals	min_match_goals	max_match_goals	iqr_match_go
top4	2.741446	3	1.694362	2.870861	0	12	
rest	2.676805	2	1.654224	2.736458	0	12	

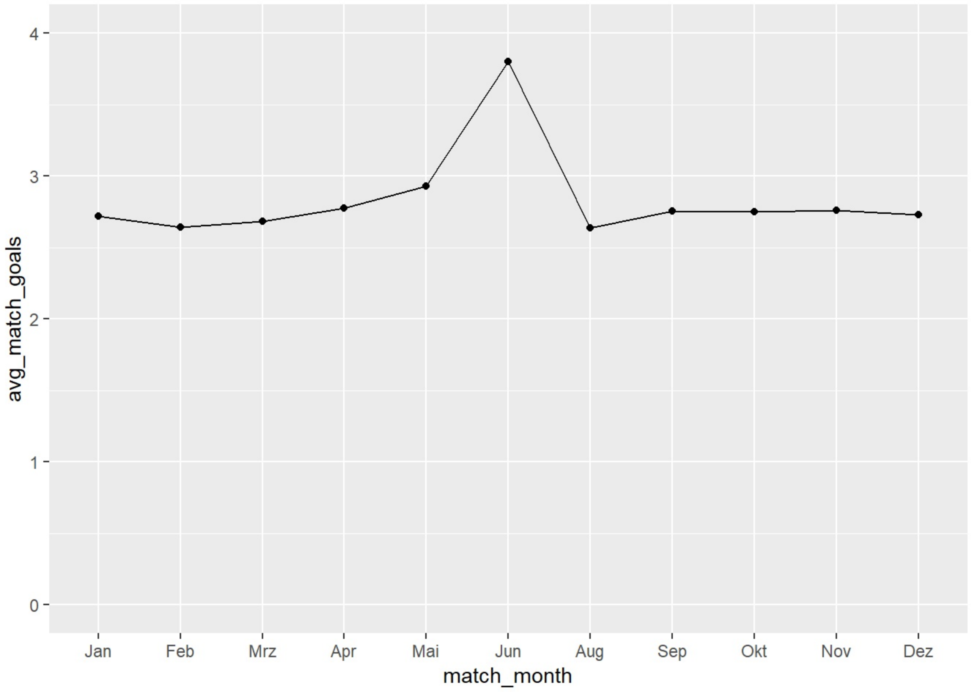
2. Is there really a home advantage? Use a box plot to show the number of goals scored by home and away teams.

```
match %>%
  gather(key = side, value = goals_scored, c(home_team_goal, away_team_goal)) %>%
  ggplot(aes(x = side, y = goals_scored)) + geom_boxplot() +
  stat_summary(geom = "point", fun.y = mean, pch = 23)
```



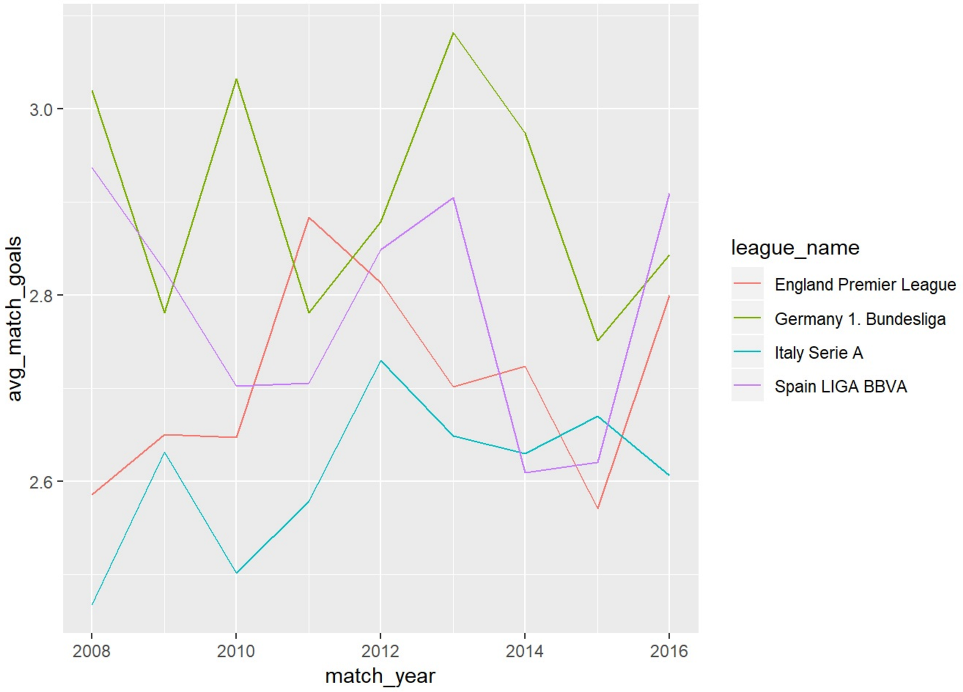
3. *"All soccer players are fair-weather players!"* Check the assertion with a line chart: Do on average more goals fall per game in the summer months than in the rest of the year?

```
match_top4 %>%
  mutate(match_month = month(as_date(date), label = T)) %>%
  group_by(match_month) %>%
  filter(!is.na(home_team_goal) | !is.na(away_team_goal)) %>%
  summarize(avg_match_goals = mean(home_team_goal + away_team_goal)) %>%
  ggplot(aes(x = match_month, y = avg_match_goals, group = 1)) +
  geom_point() +
  geom_line() +
  scale_y_continuous(limits = c(0,4))
```



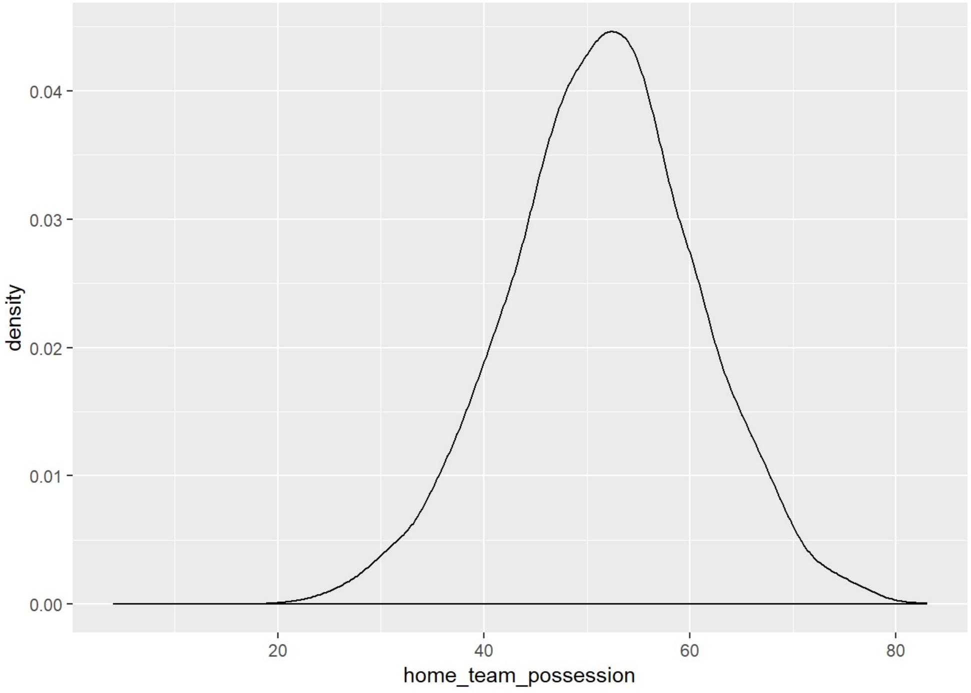
4. Display the average goals scored per game for the top 4 leagues per year from 2008 to 2016.

```
match_top4 %>%
  mutate(match_year = year(as_date(date))) %>%
  group_by(league_name, match_year) %>%
  filter(!is.na(home_team_goal) | !is.na(away_team_goal)) %>%
  summarize(avg_match_goals = mean(home_team_goal + away_team_goal)) %>%
  ggplot(aes(x = match_year, y = avg_match_goals)) +
  geom_line(aes(color = league_name))
```



5. Use an estimated density function curve AND a QQ-Plots to check whether the `home_team_possession` variable is (approximately) normally distributed.

```
# Option 1
match %>%
  ggplot(aes(x = home_team_possession)) +
  geom_density()
```

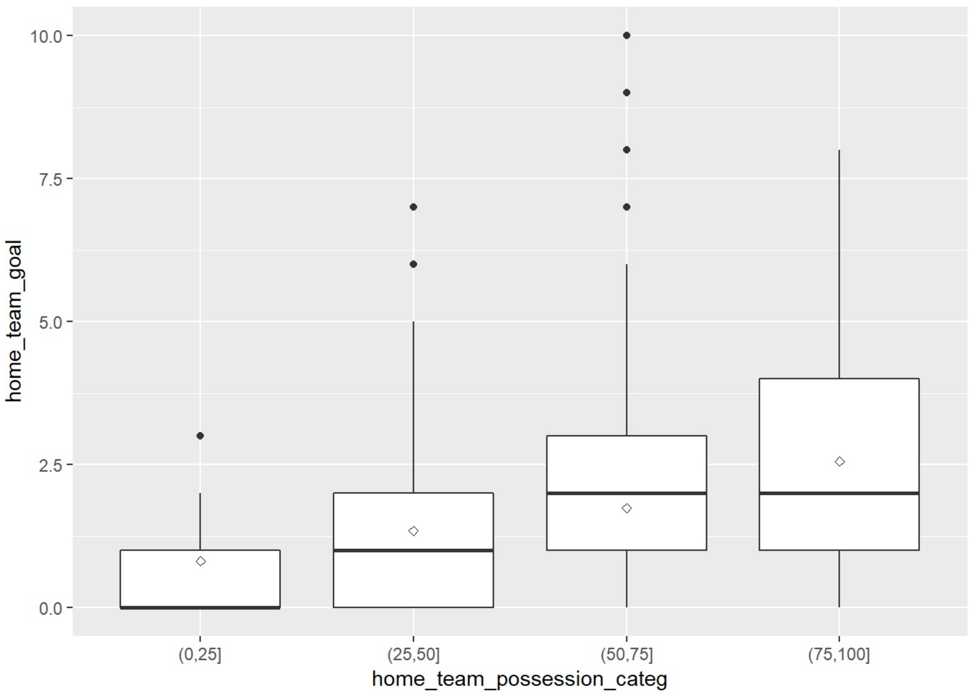


```
# Option 2
#ps <- seq(0, 100, 0.5)/100
#qs <- quantile(match$home_team_possession, ps, na.rm = T)
#normalqs <- qnorm(ps, mean = mean(match$home_team_possession, na.rm = T),
                  #sd = sd(match$home_team_possession, na.rm = T))
#plot(normalqs, qs, xlab="Normal percentiles", ylab="Data percentiles")
#abline(0,1) ##identity line

# Option 3
#qqnorm(match$home_team_possession)
```

6. Use a box plot to show whether there is a correlation between ball ownership (`home_team_possession`) and the number of goals (`home_team_goals`) scored per game for home teams. Create four categories of ball ownership shares: *very low* ($\leq 25\%$), *low* ($25\% < x \leq 50\%$), *high* ($50\% < x \leq 75\%$) und *very high* ($x > 75\%$).

```
match %>%
  filter(!is.na(home_team_possession)) %>%
  mutate(home_team_possession_categ =
           cut(home_team_possession, breaks = seq(0,100,25))) %>%
  ggplot(aes(x = home_team_possession_categ, y = home_team_goal)) +
  geom_boxplot() +
  stat_summary(geom = "point", pch = 23, fun.y = "mean")
```



Dataset:

- http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/EuropeanSoccer.sqlite (http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/EuropeanSoccer.sqlite)
  (For database schema and explanation of variables, see: https://www.kaggle.com/hugomathien/soccer (https://www.kaggle.com/hugomathien/soccer))