Aufgabenblatt 1

Die *European Soccer Database* enthält Daten zu mehr als 25.000 nationalen Fußballspielen der besten europäischen Ligen. Das Ziel dieser Übung ist, mithilfe von explorativer Datenanalyse und Visualisierung in R interessante Zusammenhänge darzustellen.

Zunächst muss auf einige der Tabellen in der Datenbank zugegriffen werden. Hinweis: Sie können dazu die Funktion `RSQLite::dbConnect()` verwenden. Um auf eine bestimmte Tabelle der Datenbank zuzugreifen und sie in ein `data.frame` umzuwandeln, können Sie den Befehl `tbl_df(dbGetQuery(connection, 'SELECT * FROM table_xyz'))` benutzen.

1. Die ersten Ligen Spaniens, Englands, Deutschlands und Italiens gelten als die vier attraktivsten Fußballligen Europas.
   a) In welcher der vier Ligen fallen im Schnitt die meisten bzw. wenigsten Tore pro Spiel?
   b) Vergleichen Sie Durchschnitt, Median, Standardabweichung, Varianz, Wertebereich (Range) und Interquartilsabstand bzgl. der pro Spiel gefallenen Tore zwischen den vier attraktivsten europäischen Ligen und den restlichen Ligen.
2. Gibt es wirklich einen Heimvorteil? Stellen Sie die Anzahl der geschossenen Tore von Heim- bzw. Auswärtsteams jeweils mithilfe eines Boxplots dar.
3. *"Alle Fußballer sind Schönwetterspieler!"* Überprüfen Sie die Behauptung mit einem Liniendiagramm: Fallen in den Sommermonaten tatsächlich durchschnittlich mehr Tore pro Spiel als im Rest des Jahres?
4. Stellen Sie die durchschnittlich pro Spiel erzielten Tore für die Top-4-Ligen pro Jahr im Verlauf von 2008 bis 2016 dar.
5. Überprüfen Sie mittels einer geschätzten Dichtefunktionskurve UND eines QQ-Plots, ob die Variable `home_team_possession` (annähernd) normalverteilt ist.
6. Zeigen Sie mithilfe eines Boxplots, ob es für Heimteams einen Zusammenhang zwischen Ballbesitz (`home_team_possession`) und Anzahl von geschossenen Toren (`home_team_goals`) pro Spiel gibt. Erstellen Sie dazu vier Kategorien von Ballbesitz-Anteilen: *sehr niedrig* ($\leq 25\%$), *niedrig* ($25\% < x \leq 50\%$), *hoch* ($50\% < x \leq 75\%$) und *sehr hoch* ($x > 75\%$).

Datensatz:

- http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/EuropeanSoccer.sqlite
  (für Datenbankschema und Variablenerklärung siehe https://www.kaggle.com/hugomathien/soccer)