Assignment 2: Data Management, Visualization, and SQL

Data Science Project Management (DS400) | Winter Term 2024/25

Formal requirements

Submit your assignment earlier than **December 10, 2024, 12:00 noon**. The submission must include your code, results, and explanations for code and results, for both **R** and **python**.

Submit your assignment via ILIAS. Upload the `*.html`-files that were rendered from your `*.Rmd` and `*.ipynb`, or `.qmd` solutions. You can optionally upload the `*.Rmd` and `*.ipynb`, or `qmd` files as well. You must use the following naming convention: `<Lastname>_<Firstname>_Assignment##_<Language>.<extension>` (e.g., `Lovelace_Ada_Assignment02_R.html`).

You will receive up to **7.5 points** upon passing this assignment.

Code of conduct

By submitting the assignment, you are acknowledging that the submitted assignment is your own work. This implies that …

- the work you submit is your own
- the code you wrote is your own
- any comment or interpretation is in your own words
- you made clear whom you worked with

Disregarding this code of conduct will result in failure of the assignment!

Data Import and Description

In this part, you will work with data from the 2024 Olympic Games in Paris. In the `data.zip` you find several csv files:

- The `athletes` dataset contains information about the athletes, for example their name, nationality, and discipline.
- The `medals` dataset contains information about each medals that was awarded, for example the discipline, the athlete, and the medal type.
- The `discipline_athletics` dataset contains information about events in the athletics discipline, for example all athletes who participated in the women's 100m run.

(1) Using your favorite IDE, set up a project so that you can use relative paths to load the data.

(2) Load the data `athletes` and `medals` data into your environment. Print the first 10 observations in each dataset. Also print all variable names for each dataset.

(3) How many different nationalities are there in the `athletes` dataset?

Data Visualization

(4) Create a barplot showing the number of athletes per nationality (from highest to lowest). What is wrong with this plot?

(5) Make another barplot where only the most common nationalities are displayed (defined as having at least 100 athletes at the Olympics). The remaining athletes should be grouped into a "other" bar. Include the actual number of athletes in each bar.

(6) Using the **medals** dataset, select the top 5 countries with the most medals. Create a plot showing the number of medals on the y-axis and the countries on the x-axis. Color the bars per gender. To make the plot easier to read, include only athletes that identify themselves as male or female. Use a colorblind-friendly palette.

(7) Re-create the previous plot but facet by medal type. Are there any differences between the medal types?

(8) Create a plot that shows the cumulative medals won by the top 3 country over time.

(9) Import the athletics dataset (**discipline_athletics.csv**). Keep only the data from women's and men's 100m and 200m finals. Print the dimensions of the resulting dataset.

(10) Create two plots (side-by-side) that shows the distribution of the athletes' ages and their finish time, once for the 100m final and once for the 200m final. Color the athletes by their gender. Do not use a shared axes for the two plots. What can you observe from the plots?

SQL

In this part, you will work with the **imdb.db** movie database. The data in this SQLite database are sampled from the IMDb database (cf. https://developer.imdb.com/non-commercial-datasets/), a database all about movies, TV series, and more, including ratings from users. The database that is provided to you contains the following tables:

- names: Contains information about people from movies, TV series etc., for example their name, birth year, and primary profession.
- titles: Contains information about the titles of the media, for example the title type, the start year, and the runtime.
- ratings: Contains information about the ratings of the titles, for example the average rating and the number of votes.

NOTE: Directly use SQL syntax for all subsequent tasks (except the creation of the histogram).

(11) Print the names of all tables in the database.

(12) Print the first 10 rows of the **names** table.

(13) How many unique names exist in the **names** table?

(14) What is the average rating of all titles in the **ratings** table?

(15) What is the average rating of all titles in the **ratings** table that have less than 100 votes?

(16) List the names, birth years, and primary professions of the 10 oldest actors/actresses in the database.

(17) Which three movies have the lowest ratings? Print their name, release year, title type, and rating.

(18) Create a histogram plot that shows the distribution of TV series ratings vs. movie ratings. What can you observe from the plot?

(19) Close the connection to the database (if applicable).