School of Computer Science

University of St.Gallen

Data Science: Exercise 5

24. October 2023

Bernhard Bermeitinger, Thomas Huber

24.10.2023

Task 3.1 - Loading additional data

- To enrich our data we will collect information about the countries. For this we will use an API.
- Make a GET request to https://restcountries.com/v3.1/all. You can use the requests library for this.
- Create a DataFrame called **countries_df** from the response
  - The response is in JSON
  - Look through the *pandas* documentation: pandas IO

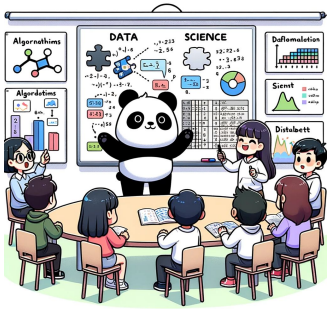
Task 3.2 - Data Cleaning

- The **name** column contains dictionaries. This makes it annoying for us to work with.
- Simplify the column by **replacing** all entries in it with the value in **common** from each dictionary per row.

Task 3.3 - Joining DataFrames

- **Combine** the two DataFrames on the **leagueCountry** column.
- For the DataFrame with the countries, you only need the **name** and **fifa** columns.

Task 4.0 - Joins



- The four most common used JOINs in SQL are:
  - **INNER JOIN**
  - **LEFT JOIN**
  - **RIGHT JOIN**
  - **FULL JOIN**
  - (there are also CROSS JOIN, self JOINs and NATURAL JOIN, and NATURAL LEFT OUTER JOIN, …)
- For these four, form one group each, familiarize yourself with what it does and come up with a real world example where it is useful.
- Break out into new groups, so at least one member of each original group is in the new groups, and discuss your JOIN in the group.

Task 4.1 - Joining with SQL

- Select all columns from the **crowdstorming** table, and only the **fifa** column from the **countries** table.
- Then join the two tables on the **leagueCountry** column of the **crowdstorming** table and the **name** column of the **countries** table.

Task 5 - Calculating the mean

- Calculate the **mean height** and **weight** of each player in the database.

Task 6 - Calculating the mean per position

- Calculate the **mean height** and **weight** of each player **per position** in the database, using SQLAlchemy.
- Repeat this, but on the **DataFrame**. Are the results the same?

Task 7 - Calculating the mean per position and league

- Calculate the **mean height** and **weight** of each player **per position** and **per league** in the database, **using** SQLAlchemy.
- Repeat this, but on the **DataFrame**. Are the results the same?

Task 8.1 - People with unusual names

- Select all people, whose first name starts with an X, from **people_database.db**.
  - You can find this file on Canvas or download via the provided snippet.
- Repeat this task, but in *pandas*:
  - Load the full database as a DataFrame.
  - Query the DataFrame for the required rows.
  - Time both: loading and querying

Task 8.2 - People with unusual names

- Select all people from **people_database.db** who share a name with a player from the **crowdstorming** table as well as the position of that player.
  Include the *fifa* column from the *countries* table as well for those players.