

Semester Project (SP)

Fitness Tracker Data Analysis

This project involves designing a MySQL database schema to store fitness tracker data, writing Python scripts to pre-process, generate and import data, and conducting exploratory data analysis (EDA) using Jupyter Notebooks. The goal is to develop a robust understanding of user activity, sleep patterns, and other health metrics provided by fitness tracker-related data.

Task Breakdown

Step 1: Database Schema Design

- Design a MySQL database schema to store data from the provided CSV files:
 - **sleep_record.csv**: Sleep records for users. Some days, users might sleep multiple times. On such days, there should only be one record, but the *TotalSleepRecord* column indicates if there were numerous sleeping sessions on the day.
 - **weight_log.csv**: Weight captures for users. Weight may be captured manually (*IsManualReport* = True) or automatically via a smart scale (*IsManualReport* = False).
 - **daily_calories.csv**: Total daily calorie intake for users.
 - **daily_activities.csv**: Records the total daily active minutes and distances for users.
 - **daily_steps.csv**: Users' total number of steps per day.
- In all of these files, the *id* field uniquely identifies individual users. The same *id* field may be found in multiple CSV files. This *id* field may be used to link activities and measurements to a specific user.
- Include an additional *Users* table with the following fields (you need to select applicable data types):
 - *id*
 - *name*
 - *surname*
 - *age*
- Write SQL `CREATE TABLE`` statements for each table in the schema.

Step 2: Data Pre-processing, Generation and Import Functions

- Write Python functions to pre-process data from each CSV file by removing rows without *id*'s or dates. In other columns with missing values, the values should be imputed using the specific column's mean. You should have one function associated with each of the CSV files.
- Write a Python function which will:
 - Process all of the CSV files and find only the unique *id*'s.

- These unique *id*'s represent the id's of the users stored in the various CSV files. You will note that the id's were provided without any data for the Users table. This is because the original data was anonymised. However, we want to have "dummy" user data to create a database. So, for each of the unique *id*'s you are required to contact a web service (any of your choosing) to generate a *name* and *surname* to associate with the *id*. For the *age*, you may randomly generate a value.
- Create a CSV file, called *users.csv*, in which each row is one of the unique *id*'s along with the generated *name*, *surname* and *age*.
- Write Python functions to import each CSV file into its corresponding table in the MySQL database. You need one function for each of the provided CSV files and one for the *users.csv* you created.

Step 3: Exploratory Data Analysis

- Use Jupyter Notebooks to perform exploratory data analysis (EDA) on the dataset. Your EDA should load data from the database and not be based on the CSV files.
- Explore each table individually and select suitable visualisations for at least some aspects of each table.
- You should also attempt to perform EDA, which crosses the various tables, to see if you can gain insights beyond those a single table can provide.
- Visualise and document at least one insight gathered from processing and exploring the data across several tables.

Deliverables

1. **SQL File:** All SQL queries for creating tables and any other database operations saved as *fitness_schema.sql*.
2. **Python File:** All Python functions for pre-processing, generating and importing data saved as *data_processing.py*.
3. **Jupyter Notebook:** The EDA process, including insights and visualisations, documented in a Jupyter Notebook named *fitness_eda.ipynb*.

Submission Instructions

- Ensure that your SQL file, Python script, and Jupyter Notebook are named as specified above.
- Zip these three files into a single archive named *emailaddress.zip*. The email address part of the file name should be your Noroff email address.
- Submit this zip file as your final submission for the project.

Begin your project by structuring your approach to each component, ensuring that you meet all the requirements stipulated for the database design, data pre-processing, generation and importation, and exploratory data analysis. Good luck!