Friedrich-Alexander-Universität
Technische Fakultät

Department of Computer Science
Computer Science 6
(Data Management)

Knowledge Discovery in Databases with Exercises
Summer Semester 2024

Exercise Sheet 2:
Data Analysis and Preprocessing

About this Exercise Sheet

This exercise sheet focuses on the content of the lectures *3. Getting to Know Your Data* (Exercise 1) and *4. Data Preprocessing* (Exercises 2 and 3). The goal of this exercise sheet is to familiarize students with the content through a practical data science example.

The exercise sheet is designed to span three weeks and will thus be covered in three exercise sessions. The plan is to tackle one exercise per session. However, if there is not enough time for an exercise in any session, it can be continued in the following session. Similarly, if there is extra time, work can commence on the next exercise.

To accommodate this flexible model, the sample solution will be published only after the three weeks have elapsed.

Preparation

Before participating in the exercise, you must prepare the following:

1. **Install Python and pip on your computer**
   - Detailed instructions can be found in `1-Introduction-Python-Pandas.pdf`.
2. **Download provided additional files**
   - Download `Additional-Files-Student.zip` from StudOn
   - Extract it to a folder of your choice.
3. **Install required Python packages**
   - Open a terminal and navigate to the folder where you extracted the files.
   - Run the command `pip install -r requirements.txt` within the extracted additional files folder to install the required Python packages.

Exercise 1: Getting to Know Your Data

This exercise comprises practical data science tasks and thus utilizes a Jupyter Notebook:

1. Open `Getting-to-Know-Your-Data.ipynb`.
2. Take a look at the tasks (blue boxes) in the notebook and try to solve them.

If you are unfamiliar with how to open a Jupyter Notebook, please refer to Exercise 1 of `1-Introduction-Python-Pandas.pdf`.

If you have worked through `1-Introduction-Python-Pandas.pdf` some of the initial tasks may seem familiar to you. However, this is a good repetition to consolidate the methods that you will need more often during the semester.

The solution to the exercise can be found in `Additional-Files-Solution.zip`.

Exercise 2: Data Cleaning and Integration

This exercise comprises practical data science tasks and thus utilizes a Jupyter Notebook:

1. Open `Data-Cleaning-and-Integration.ipynb`.
2. Take a look at the tasks (blue boxes) in the notebook and try to solve them.

If you are unfamiliar with how to open a Jupyter Notebook, please refer to Exercise 1 of `1-Introduction-Python-Pandas.pdf`.

The solution to the exercise can be found in `Additional-Files-Solution.zip`.

Exercise 3: Data Reduction, Transformation, and Discretization

This exercise comprises practical data science tasks and thus utilizes a Jupyter Notebook:

1. Open `Data-Reduction-Transformation-and-Discretization.ipynb`.
2. Take a look at the tasks (blue boxes) in the notebook and try to solve them.

If you are unfamiliar with how to open a Jupyter Notebook, please refer to Exercise 1 of `1-Introduction-Python-Pandas.pdf`.

The solution to the exercise can be found in `Additional-Files-Solution.zip`.