Big Data

Management and Analytics

Assignment 1

TU Clausthal, Institut für Informatik

!!! Due date: 14 May 2025, 1pm!!!

In this assignment, you will further practise your programming skills with Python and get some first experience with the Python library Pandas (`https://pandas.pydata.org/`) for big data analysis.

Please ensure that Python (version 3.9 or higher) is set up and running for future tutorials.

Read the following notes in detail for your submission of solutions.

- You should work in teams of up to 2
- Please pay attention to the deadline.
- Copying solutions from other students will be treated as cheating and will lead to exclusion from the course!
- If you used any sources or tools to complete the tasks, you must clearly specify both the source and the reason for its use. Failure to do so may be considered plagiarism.
- Document your solution clearly: Include code and queries in both a .txt file and a .pdf file containing your explanation, which may be used for presentation; use proper tools for diagrams, and cite all sources to avoid plagiarism.
- You can send us your solutions via the **Moodle**

Here are some useful links for learning more about Python:

Mark Pilgrim - Dive into Python: `http://www.diveintopython3.net/`

Python 3.9 Documentation: `https://docs.python.org/3.9/`

!! Submit your solutions via Moodle !!

Task 1 **5 Marks**

In the first task, we use so-called *lambda-functions*. You can get general information about it if you follow this link: `https://www.w3schools.com/python/python_lambda.asp`

(Important note: Do NOT mix them up with the term "lambda architecture"!)

We use the map, filter, and reduce paradigms. For them, you can find more information and tutorials here: `https://www.learnpython.org/en/Map%2C_Filter%2C_Reduce`. Solve the following tasks and write your solutions into a .txt file.

(a) Write a list comprehension that takes a number n and returns a list of even numbers, using a lambda function.

(b) First, write a function that takes a length in inches and returns a length in cm. Given a list l with lengths in inches: $l = [4, 4.5, 5, 5.5; 6, 7]$. Write a list comprehension that takes l and returns a list with all values converted to cm using *map*.

(c) Write a list comprehension, which filters the list l from (b) by returning only sizes between 4 and 6 inches. Use *filter* for this!

(d) Write a list comprehension that reduces the list l by summing up all lengths.

Hint: For using the *reduce* function, you need to import it first by adding the line: *from functools import reduce*

Task 2 **5 Marks**

In this task, we use the *Pandas* library for big data analysis. For the documentation, see here: `https://pandas.pydata.org/pandas-docs/stable/`. We use data records of a movie database, which is provided as a CSV file in Moodle, namely "*Moviedata.csv*". Do the following tasks and write the code that solves the task into a .txt file.

(a) Read the .csv-file as a DataFrame for further processing using *pandas.read_csv()*. Afterwards inspect the read .csv-file using *.shape*, *.columns*, *.info* and *.describe()*. Lastly, display the first five records of the dataset using *.head(5)* and the last five records using *.tail(5)*.

(b) Select the first five records from the data set, but those records shall only contain the following columns in your output: *movie title*, *duration* and *num voted users*. Write the code that achieves this output.

(c) Select the first five movies containing the genre "*Action*". Display only the columns "*movie*", "*title*" and "*genres*".

(d) Sort the action movies by their *IMDb score* and display the names and scores of the top-10 scored movies.

(e) Group the movies by column *director* and display the top-10 directors with the highest mean *gross* of their movies.

(f) Delete all rows that contain at least one missing value. Visualize parts of the data using *pandas:plotting:scatter_matrix* and *DataFrameGroupBy:hist*.

+++++