Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

The second part of the Assignment, IDS 2020-2021

Introduction

The assignment guides you through the analysis of several datasets using techniques and tools provided in the course. The second part of the assignment tests your understanding of the material discussed in the lectures 10-21. The assignment questions are given in the Jupyter notebook. It is necessary to follow the assignment in the given order because certain questions might require an output obtained in the previous steps. Please note that it is important to use **the Python environment provided for this course** to answer the questions.

The datasets

In contrast to the first assignment, we use different datasets for different questions in this assignment. Here you can find a brief description of each dataset.

DataPrepViz (Questions 1 and 2)

This real-life dataset from the year 2015 contains information about several properties of different countries in different parts of the world. There are the following seven features:

- *country*: The name of the country.
- *geographic_group*: Each country is mapped to a certain group based on geographic location.
- *children_per_woman_total_fertility*: This fertility metric provides the average number of children per woman
- *child_mortality_0_5_year_olds_dying_per_1000_born*: The number of children dying at the age between 0 and 5 years, per 1000 children born.
- *co2_emissions_tonnes_per_person*: The CO2 emissions of the country in tonnes per person per year.

- *corruption_perception_index_cpi*: The corruption perception score. The higher the score, the higher the corruption perceived by the residents in average.
- *life_expectancy_years*: The number of years an average person is expected to live.
- *vccin_effect_dag*: The vaccine confidence score. Higher numbers indicate a stronger belief in the effect and importance of vaccination.

Store_data (Question 3)

It is a real-life dataset containing information about transactions over the course of a week at a French retail store. Each row shows a transaction containing items such as egg, milk, etc. Since this dataset is real, the diversity of behaviors in that is high.

Sms_data (Question 4)

The dataset is a public set of labeled SMS messages that have been collected for mobile phone spam research. It is a collection composed of 5.574 English, real and non-encoded messages, tagged according to being legitimate (ham) or spam. More about the dataset can be found at http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection.

Quarantine_Log (Questions 5 and 6)

This event log is created with CPN tools (based on a model mainly created by Tobias Brockhoff) simulating the flow of corona quarantine cases in Germany. It is an artificial event log inspired by the real rules (in summer 2020).

If mentioned in the question, perform the data preprocessing as explained in the task in the Jupyter notebook and export the resulting sampled dataset. Submit all the sampled datasets with your results.

Submission and deliverables

The deadline for the assignment is 12/02/2021 23:59. You will need to hand in your submission via Moodle. Please note, that a deadline extension is not possible and late submissions will not be considered.

The assignment should be done in groups of 2-3 students. It is not mandatory that the group members remain the same as the ones for the first part of the assignment. Make sure to include all group member names and their ids in the jupyter notebook.

Your submission should include a **Jupyter notebook** with your results and your Python code that indicates how you have obtained the results. In addition, please, upload a **zip-file** with all requested datasets and other outputs, in particular, the sampled datasets created in the preprocessing steps. An additional report is **NOT** required and will not be considered for grading.

Submission summary:

You have to submit two items to the Moodle:

1. **A Jupyter notebook.**
   - Use the provided Jupyter notebook to present your results and code.
   - Make sure that names and student IDs of all group members are provided in the notebook.
2. A '**datasets.zip**' with all requested datasets and other outputs, such as pdf, jpg, etc.. Please, do not forget to discuss the outputs in the notebook.

Grading

Successful participation in the assignment, i.e. scoring at least 50%, is one of the prerequisites for taking the written exam. The results of the assignment are valid only in the current semester and will expire afterwards. The assignment can only be redone in the next academic year.

The assignment counts as 40% of the final grade (20% each part). In the second part of the assignment, 100 points are assigned: 90 points for the main sections and 10 points related to your reporting style:

1. Data Preprocessing and Data quality – 15 points
2. Data Preprocessing and Advances Visualization – 15 points
3. Frequent item Sets and Association Rules – 15 points
4. Text Mining – 15 points
5. Process Mining – 15 points
6. Big Data – 15 points

For a data scientist, a sufficient and proper presentation of the results is as much important as analysis and code. Therefore, 10 points are given for your reporting style. A few useful hints on how to present your work well:

- Add comments to your code.
- Do not mix code and answers to the questions. Use markdown cells to explain your solutions and provide answers to the stated questions in a sufficient and coherent manner.
- Do not change the structure of the notebook, except of adding blocks for better readability.
- In general, make your answers readable and understandable. Please, check the spelling.

Please note, that correct and full results, sound code and sufficient explanations are highly important.