Data Validation Efficiency Framework

Context:

A lot of issues in implemented data pipelines can affect the quality of the processed data (e.g. data integration and transformation, over- & underflow errors). Currently, there are a variety of different data validation techniques (e.g. correlation analysis, hypothesis testing, uniqueness constraints etc.) to detect such issues. However, currently there is no overview which data validation techniques can detect which data quality issues.

Goal:

Development of a framework that assesses the efficiency of current data validation techniques according to the most occurring data processing errors in typical data pipelines.

Procedure:

- Multivocal Literature review (overview of current data validation techniques)[1]
- Informal Literature review on most common data processing issues in data pipelines[2]
- Analysis of efficiency of data validation methods according to data processing issues (is the technique able to detect the issue, only in specific cases etc. (**benefit**); effort for implementing and maintaining the technique (**costs**))
- Based on results of the analysis a framework should be developed that provides an overview and supports further analysis
- Evaluation of the framework based on a survey, expert interviews or a case study

There is already a certain amount of preparatory work and literature available on which to build.

[1] Data Validation Libraries:

https://docs.greatexpectations.io/en/latest/reference/expectation_glossary.html
https://pypi.org/project/voluptuous/
https://engarde.readthedocs.io/en/latest/
https://www.tensorflow.org/tfx/guide/tfdv

[2] Data processing errors:

https://peda.net/kenya/css/subjects/computer-studies/form-three/driac2/data-processing/doeidp