1 Sample Project - **Data Report**

All information on the data used in the project is compiled in the data report in order to ensure the traceability and reproducibility of the results and to enable a systematic expansion of the database.

Typically, in the exploratory analysis of the acquired raw data, quality and other issues are identified, which require pre-processing, merging of individual datasets and feature engineering into processed datasets. Therefore, this template provides a separate section for the processed data, which then serves as a starting point for the modelling activities. This needs to be adapted to the specific project requirements.

1.1 Raw data

1.1.1 Overview Raw Datasets

Name	Quelle	Storage location
Dataset 1	Name/short description of the data source	Link and/or short description of the location where the data is stored, e.g. accessible to the team
Dataset 2	...	...

1.1.2 Details Dataset 1

- Description of what information the dataset contains
- Details of the data source/provider
- Information on data procurement: description and possibly references to resources (download scripts, tools, online services, …). Any new team member should be able to acquire the data indepentendently following these instructions.
- Legal aspects of data use, licences, etc.
- Data governance aspects: Categorisation of the data based on internal business requirements, e.g. public, business-relevant, personal
- If applicable: categorisation into dependent (target variable, regressor) and independent (regressor) variables

- …

1.1.2.1 Data Catalogue

The data catalogue basically represents an extended schema of a relational database.

Column index	Column name	Datatype	Values (Range, validation rules)	Short description
1				
2				

1.1.2.2 If applicable: Entity Relationship Diagram

1.1.2.3 Data Quality

1.2 Processed Data

1.2.1 Overview Processed Datasets

Name	Quelle	Storage location
Processed Dataset 1	Name/short description of the data source	Link and/or short description of the location where the data is stored, e.g. accessible to the team
Processed Dataset 2	…	…

1.2.2 Details Processed Dataset 1

- Description of what information the dataset contains
- Details and reasons for the processing steps -> Traceability and ensuring reproducibility
- How can the data be accessed? Description, scripts, tools, …
- …

1.2.2.1 Data Catalogue

1.2.2.2 If applicable: Entity Relationship Diagram

1.2.3 Details Processed Dataset 2

…