Tool: Data Profile

The Data Profile is an exploratory analysis that provides information to evaluate the quality, integrity, temporality, consistency, and possible biases of a dataset that will be used to train a machine learning model (Gebru et al. 2018).

Data and Input Collection and Origin	
Name of dataset used.	
What institution created the dataset?	
For what purpose did the institution create the dataset used?	
What mechanisms or procedures were used to collect the data (e.g., household survey, sensor, software, API)? Does they comply with existing data protection regulations?	
What is the scale of the dataset?	
Obtain documentation for each variable within the dataset. Provide a short description, including its name and type, what it represents, how it is measured, etc.	

Data and Input Domains

What is the data domain (e.g., proprietary, public, personal)?	
If personal, are the data identified, pseudonymized, unlinked pseudonymized, anonymized, or aggregated?	
If proprietary, are any intellectual property rights considerations?	

Data and Input Structure

Are the data static or dynamic? If dynamic, how often will they be updated?	

Data Quality & Qualification

How were the data obtained (observed, derived, synthetic, or provided by individuals or organizations)?	
Is the data representative of the population of interest?	
Analyze spatial and temporal coverage of the data.	
Analyze coverage of protected groups (sex, race, age, etc.).	
Describe the type of sampling used to obtain the data.	

Describe the important dimensions in which the data sample may differ from the population, in particular unmeasured selection biases. Use literature related to the subject and information from experts.	
Identify possible "undesirable states" in the data that could lead to prejudicial biases and inequities for a given subgroup, or any other pattern that is considered suboptimal or undesirable from a social policy point of view.	
Are there any missing values? If so, explain the reasons why this information is not available (this includes information intentionally removed). Identify reasons	