# Data methodology and observability

Charles T. Gray

September 20, 2025

## Contents



**Abstract**

Validation and observability of prioritisation of UN SDGs represented in two surveys facilitated and collected by Garima Gupta. Participants in one survey from the North Indian regions Tehri and the other survey participants from Arunachal Pradesh.

This document outlines the data methodology of the github repository and observability framework for data governance.

This project employs a structured approach to data management, ensuring data quality, integrity, and usability through various layers of data observability. The observability framework is designed to monitor and validate data at each stage of the pipeline, providing transparency, accountability, and extensiblity in the data handling processes.

This document and the associated data are curated for submission to the beta rollout of the Frontiers FAIR data management platform.

# 1 What are these data?

The data in this project aggregates two region-based surveys, with responses from participants in Tehri and Arunachal Pradesh, focusing on their ranked top three priorities of the United Nations Sustainable Development Goals (UN SDGs). The data was collected by Garima Gupta by surveying residents of the two regions in Hindi. The data is stored in the `analysis_data` directory. This project was created by Charles T. Gray to aggregate survey responses across regions to investigate the following question.

> *What are the differences in SDG priorities between residents of Tehri, where a dam was constructed 20 years ago for hydroelectric power, and Arunachal Pradesh, where a dam is currently being developed?*

Figure 1 shows a treemap of the top 3 UN SDG priorities from survey respondents in Tehri (dam constructed 20 years ago) and Arunachal Pradesh (dam under development), North India.

By structuring the data in tidy format (one row per ranking by one respondent),

# 2 This data is FAIR

## 2.1 Findable

CSV files respresenting respondent metadata and top 3 priorities of UN SDGs are found in `analysis_data`. The data are described in the `data_dictionary.md` file. The data are version controlled in the github repository `sig_gg`.

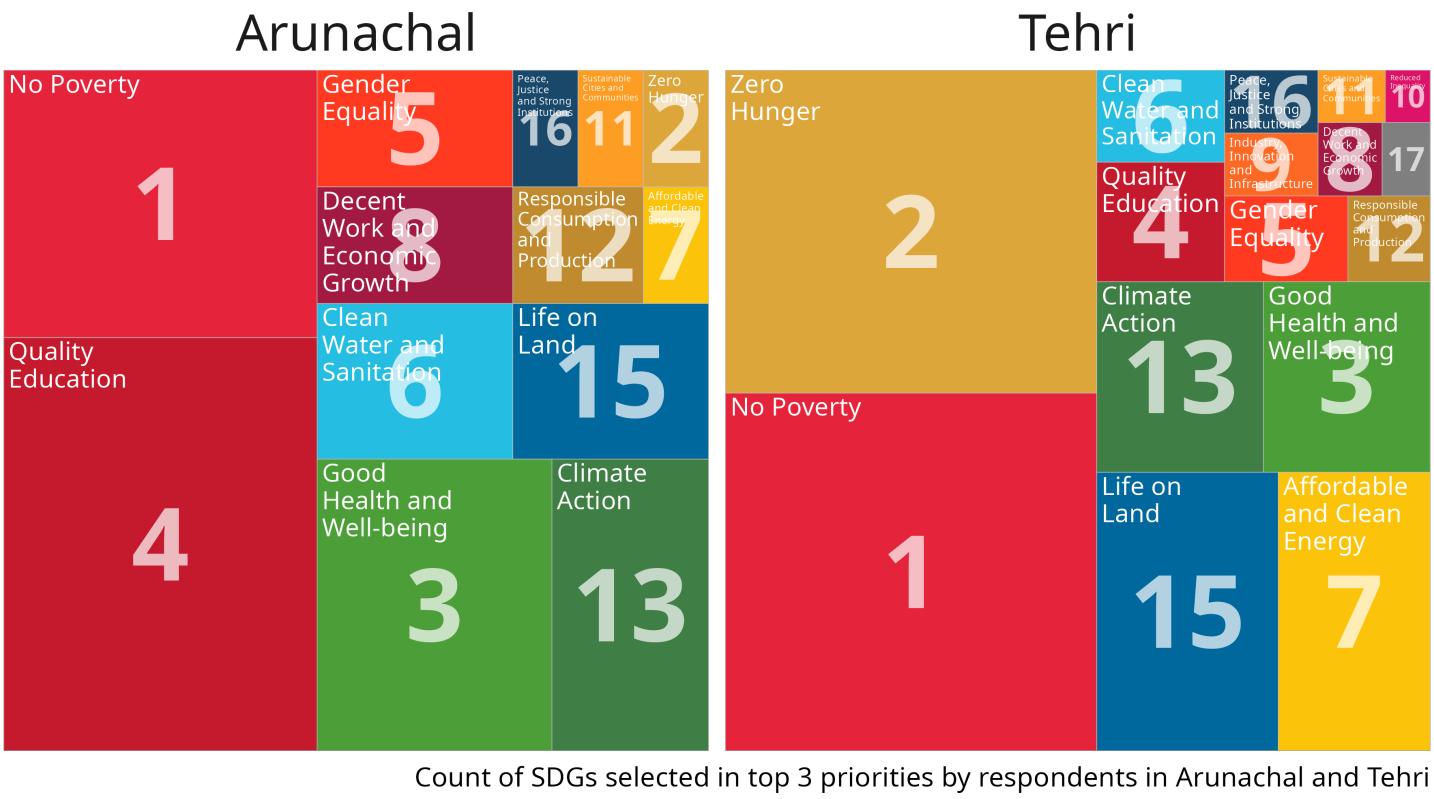


Figure 1: Treemap of top 3 UN SDG priorities from survey respondents in Tehri (dam constructed 20 years ago) and Arunachal Pradesh (dam under development), North India. Each rectangle represents a specific UN SDG, with its size proportional to the number of respondents who ranked it among their top three priorities. The treemap visually highlights the most and least prioritized SDGs in each region, providing insights into regional differences in development priorities.

## 2.2 Accessible

The data are provided in tidy format with descriptions, columns, and validations detailed in DBT. Column names were chosen for interpreability.

### 2.2.1 Responses to top 3 priorities of UN SDGs

These responses are enriched with respondent metadata that corresponds to the respondent dataset also provided.

- **response_id:** Unique identifier for each response.
- **respondent_id:** Unique identifier for each respondent
- **region:** Region where the data was collected (Tehri or Arunachal Pradesh).
- **sdg_rank:** Rank of the UN SDG priority (1, 2, or 3).
- **sdg_number:** Unique identifier for each UN SDG (1-17).
- **sdg_label:** UN SDG as text label.
- **age:** Age of the respondent.
- **gender:** Gender of the respondent.
- **displacement_status:** Displacement status of the respondent.

By enriching the responses with respondent metadata, we can analyse responses in the context of respondent demographics and characteristics. For example, in Figure 2, we can see the distribution of top 3 UN SDG priorities by gender.

#### 2.2.2 Respondent metadata

- **respondent_id:** Unique identifier for each respondent.
- **region:** Region where the data was collected (Tehri or Arunachal Pradesh).
- **age:** Age of the respondent.
- **gender:** Gender of the respondent.
- **displacement_status:** Displacement status of the respondent.

### 2.3 Interoperable

The data are published in `.csv` format on a publically available github repository `sig_gg`, readable by url into Python or R.

### 2.4 Reusable

The data pipeline is extensible. The repository can be forked and the DBT pipeline extended to transform the data in additional ways or patch in additional source data. The data are licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, allowing others to share and adapt the data with appropriate credit.

## 3 Data lineage

In `raw_data` the original two survey data are found. To aggregate across the `.csv` and Kobo tool data export in `.xlsx` format, a data transformation pipeline was implemented using the Data Build Tool (DBT) in the `dbt_project` directory. The DBT project consists of SQL scripts, tests, and documentation that transform the raw data into a tidy format suitable for analysis. Identifying which columns to extract from the Kobo export was nontrivial, `scripts/tehri-cols.R` contains the script used to identify the relevant columns, and a markdown table identifying the columns associated with survey questions can be found in `analysis_data/tehri.md`.

Figure 3 shows the directed acyclic graph (DAG) for the Data Build Tool (DBT) project `dbt_project` that produced the .csv data found in `analysis_data`. The DAG illustrates the relationships and dependencies between various data models, tests, and sources within the project. Each node represents a table produced during transformation, while the edges indicate the flow of data and dependencies among them. This visualization illustrates how data is transformed and validated throughout the pipeline, ensuring data integrity and reliability.

Table 1 lists the DBT models created by the `dbt_project`, ordered by observability layer.

Table 1: Data Build Tool (DBT) models created by dbt_project, ordered by observability layer.

| Model | Description |
|---|---|
| **Analytic Models** | |
| ana_respondents | NA |
| ana_top3 | NA |
| **Semantic Models** | |
| sem_respondents | NA |
| sem_sdg_labels | NA |
| sem_top3 | NA |
| **Source Entity Models** | |
| se_respondents_arunachal | Each row represents a respondent from Arunachal Pradesh. |
| se_respondents_tehri | Each row represents a respondent from Tehri. |
| se_top3_arunachal | Each row represents a respondent's ranking of their top 3 UN Sustainable Development Goals (SDGs) for Arunachal Pradesh, India. |
| se_top3_tehri | Each row represents a respondent's ranking of their top 3 UN Sustainable Development Goals (SDGs) for Tehri, India. |
| **Base Models** | |
| base_arunachal | Raw data from survey in Arunachal Pradesh, collected by Dr Garima Gupta. Each row represents the survey responses of a single respondent. |
| base_tehri | Raw data from survey in Tehri, collected by Dr Garima Gupta. Each row represents the survey responses of a single respondent. |

# 4 Data observability

## 4.1 Layers

The data observability framework implemented in the `dbt_project` is structured into four distinct layers: base models, source entities, semantic models, and analytic models. Each layer serves a specific purpose in ensuring data quality, integrity, and usability throughout the data pipeline.

- **Base Models:** These are the foundational models that directly interact with raw data sources. They perform initial data extraction and loading (ETL) processes to prepare the data for further analysis. Base models focus on cleaning and structuring the data to ensure it is in a usable format. These models were loaded from CSV files as seeds. The script to convert the raw files to csv to be seeded is found in `scripts/seed-data.R`.
- **Source Entities:** This layer represents the core data entities (responses and respondents) by source derived from the base models for respondents and responses. This layer allows for easier validation of data integrity as each source has different labelling and structure.
- **Semantic Models:** Semantic models aggregate source entities to semantic entities: respondents, responses. This layer ensures the consistency of labelling, for example, UN SDGs were labelled very differently in the two datasets.
- **Analytic Models:** The analytic layer enriched semantic data. For example, responses are enriched with respondent metadata such as region, age, gender, and displacement status.

## 4.2 Data validation

Table 2 lists the DBT tests applied to `dbt_project` transformations, ordered by observability layer. These tests tell us what assumptions we have validated at each transformation step in the lineage (Figure 3).

Table 2: Data Build Tool (DBT) tests applied to dbt_project models, ordered by observability layer.

| Model | Test | Columns | Arguments | Result |
|---|---|---|---|---|
| **Analytic Model Tests** | | | | |
| ana_respondents | unique_combination_of_columns | id_respondent | NA | success |
| **Semantic Model Tests** | | | | |
| **Source Entity Tests** | | | | |
| se_respondents_arunachal | accepted_values | gender | Male__Female__Prefer_not_to_say | success |
| se_respondents_arunachal | not_null | id_respondent | NA | success |
| se_respondents_arunachal | unique | id_respondent | NA | success |
| se_respondents_arunachal | unique_combination_of_columns | id_respondent | NA | success |
| se_respondents_tehri | accepted_values | gender | Male__Female__Prefer_not_to_say | success |
| se_respondents_tehri | not_null | id_respondent | NA | success |
| se_respondents_tehri | unique | id_respondent | NA | success |
| se_respondents_tehri | unique_combination_of_columns | id_respondent | NA | success |
| se_top3_arunachal | accepted_values | rank | 1__2__3 | success |
| se_top3_arunachal | not_null | id_respondent | NA | success |
| se_top3_arunachal | unique_combination_of_columns | id_respondent | rank | success |
| se_top3_tehri | accepted_values | rank | 1__2__3 | success |
| se_top3_tehri | not_null | id_respondent | NA | success |
| se_top3_tehri | unique_combination_of_columns | id_respondent | rank | success |
| **Base Model Tests** | | | | |
| base_arunachal | unique_combination_of_columns | id_respondent | NA | success |
| base_arunachal | unique_combination_of_columns | respondents | NA | success |
| base_tehri | not_null | id_respondent | NA | success |
| base_tehri | unique | id_respondent | NA | success |
| base_tehri | unique_combination_of_columns | id | NA | success |
| base_tehri | unique_combination_of_columns | id_respondent | NA | success |

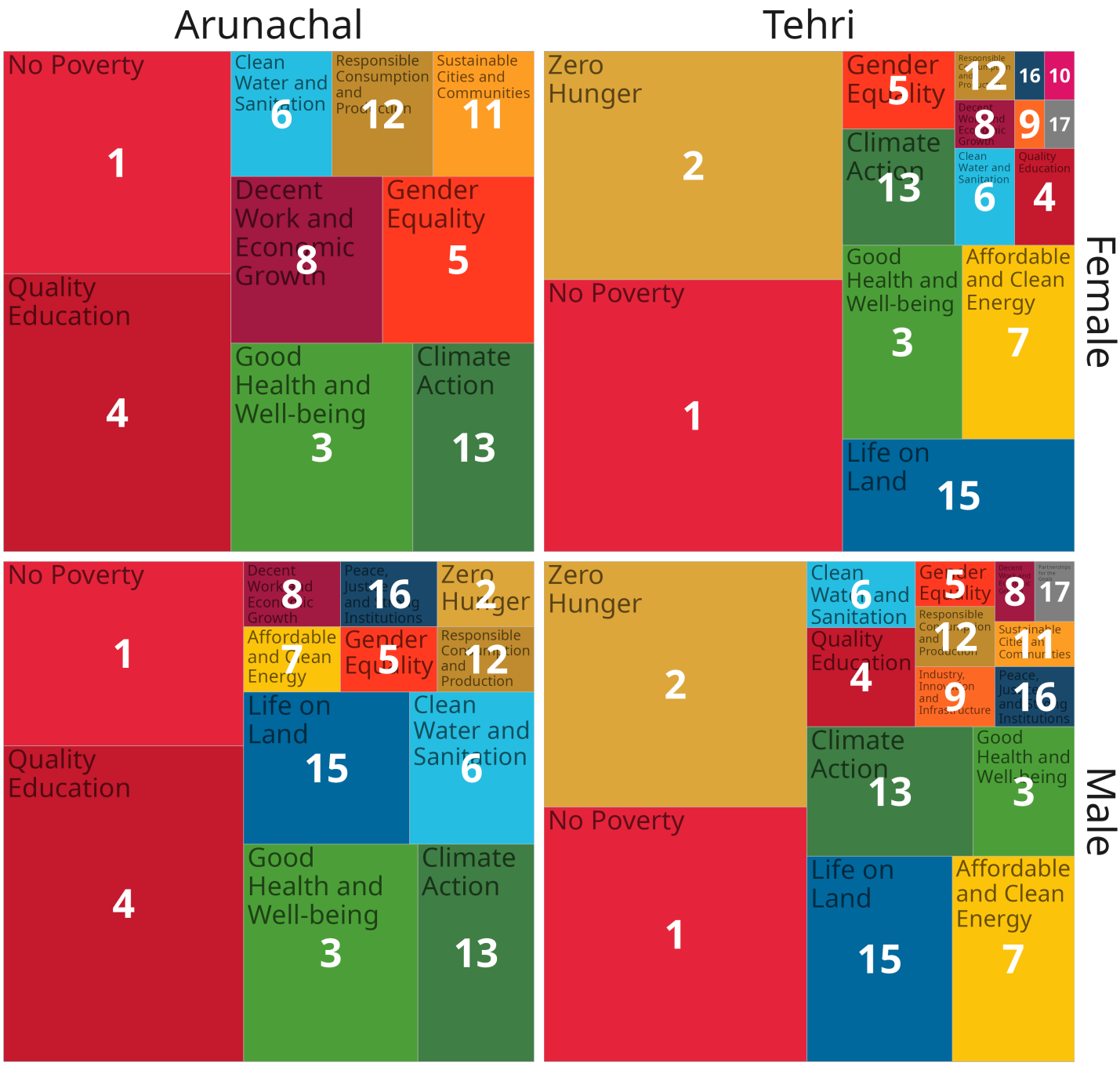


Figure 2: Treemap of top 3 UN SDG priorities from survey respondents in Tehri and Arunachal Pradesh, North India, by gender. Each rectangle represents a specific UN SDG, with its size proportional to the number of respondents of a particular gender who ranked it among their top three priorities. The treemap visually highlights the most and least prioritized SDGs in each region, providing insights into gender differences in development priorities.

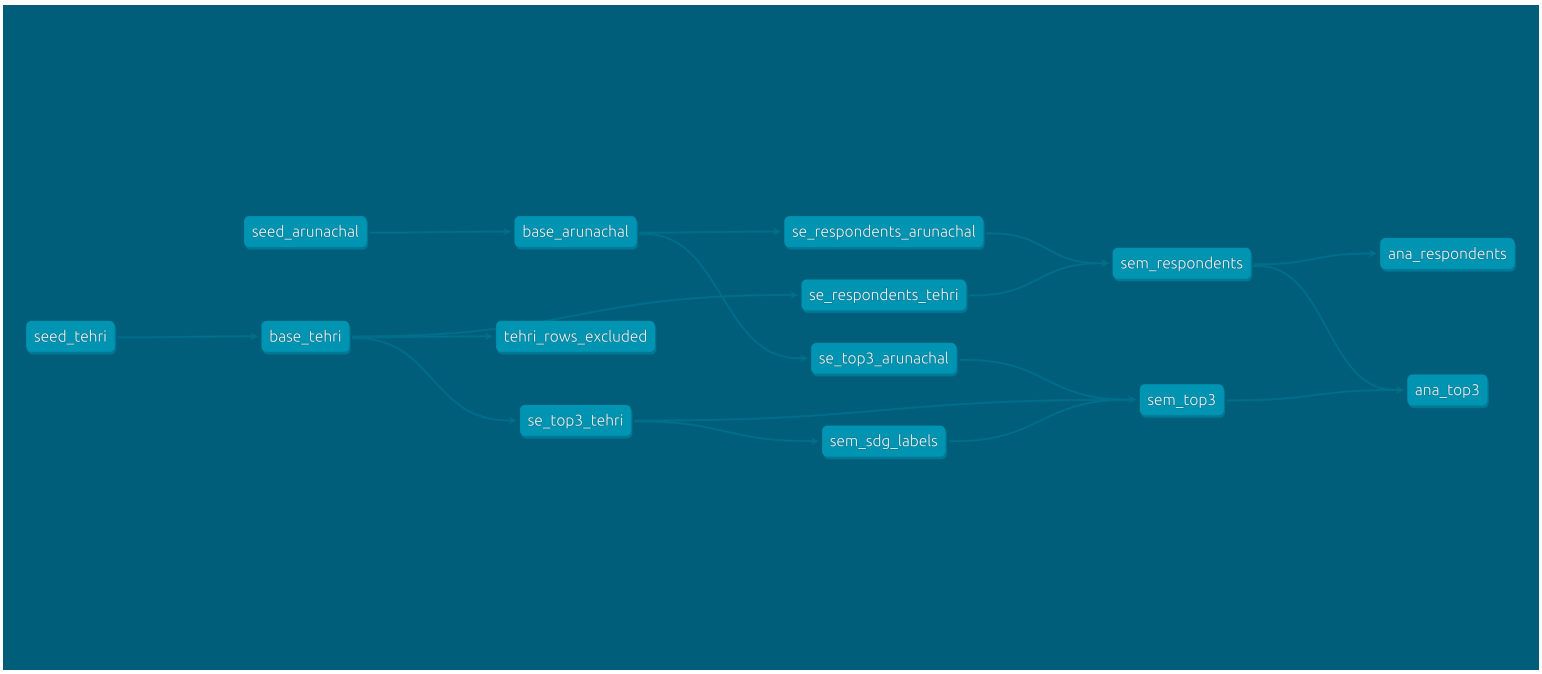


Figure 3: Data Build Tool (DBT) directed acyclic graph (DAG).