# SAMS Data Documentation

This document serves as a reference for all cleaned data files generated by the SAMS preprocessing pipeline.

## File Format

All files are stored in **Parquet (.pq)** format for efficient storage and fast reads. Parquet files can be opened using:

- Python: pandas.read\_parquet() or polars.read\_parquet()

- R: arrow::read\_parquet()

- Stata: We don’t know. If you must use Stata, you should convert to a csv, xlsx, or dta format and then read it in.

## ITI (Industrial Training Institutes) Data Files (located in path Cleaned/ITI)

### Student-Level Data

#### 1. **iti\_enrollments.pq**

**Description**: Student level enrollment data for ITI programs

**Key Fields**:

- barcode: Unique student identifier

- academic\_year: Year of application (2017-2024)

- gender: Student gender

- social\_category: Caste category (GEN, OBC/SEBC, SC, ST)

- state, district: Location information –

course\_period: Duration of the course

- module: Always “ITI”

#### 2. **iti\_marks.pq**

**Description**: Student level data containing fields describing academic performance in the class 10 exams for students enrolled in ITI programs

**Key Fields**:

- academic\_year: Year of application

- total\_marks: Maximum possible marks

- secured\_marks: Marks obtained by student

- percentage: Marks obtained as a percentage of marks possible

- compartmental\_fail\_mark: Pass and fail marks for compartment subjects - subject\_wise\_marks: Marks obtained per subject with maximum marks

#### 3. **iti\_applications.pq**

**Description**: Application data for it is.

**Key Fields**: - barcode: Unique student identifier  
- academic\_year: Year of application  
- annual\_income: Annual income of the student’s household  
- social\_category: Caste category

- institute\_name: Name of the institute applied to

- reported\_institute: Institute where student reported

- trade: Course information  
- reported\_branch\_or\_trade: Branch or trade where the student enrolled

- phase: Admission round number

- option\_no: Student’s preference order for the institute/trade (1 = first choice, 2 = second, etc.)  
- num\_applications: Total number of applications submitted by the student

### Institute-Level Data

#### 3. **iti\_institutes\_strength.pq**

**Description**: Institute level data on sanctioned strength and seat allocation for ITI programs by trade

**Key Fields**:

- sams\_code: Unique institute identifier

- academic\_year: Year of application

- strength: Total available seats

- trade: Course information (for ITI programs)

#### 4. **iti\_institutes\_cutoffs.pq**

**Description**: Institute level data on cutoff marks for ITI institutes **Key Fields**:

- sams\_code: Unique institute identifier

- academic\_year: Year of application

- trade: Course information

- cutoff: Minimum marks required for admission in the previous academic year

- social\_category: Student eligibility category

- institute\_name: Name of the institute

#### 5. **iti\_institutes\_enrollments.pq**

**Description**: Institute level enrollment data for ITI institutes

**Key Fields**:

- sams\_code: Unique institute identifier

- academic\_year: Year of application

- enrollments: Number of enrolled students

- institute\_name: Name of the institute

- category: Reservation category

- module: Always “ITI”

## Diploma / Polytechnic Data Files (located in path (Cleaned/Diploma))

### Student-Level Data

#### 6. **diploma\_enrollments.pq**

**Description**: Student level enrollment data for Polytechnic / Diploma programs

**Key Fields**:

- barcode: Unique student identifier

- academic\_year: Year of application (2018-2024)

- gender: Student gender

- social\_category: Caste category (SC, ST, Other)

- state, district: Location information

- module: Always “Diploma”

#### 7. **diploma\_marks.pq**

**Description**: Student level data containing fields describing academic performance in the class 10 exams for students enrolled in Polytechnic / Diploma programs

**Key Fields**: - academic\_year: Year of application - year\_of\_passing: Year in which the highest qualification was completed  
- total\_marks: Maximum possible marks  
- secured\_marks: Marks obtained by student  
- percentage: Marks obtained as a percentage of marks possible  
- compartmental\_fail\_mark: Pass and fail marks for compartment subjects  
- subject\_wise\_marks: Marks obtained per subject with maximum marks

### Institute-Level Data

#### 8. **diploma\_institutes\_strength.pq**

**Description**: Institute level data on sanctioned strength and seat allocation for Polytechnic / Diploma programs by branch

**Key Fields**:

- sams\_code: Unique institute identifier

- academic\_year: Year of application

- strength: Total available seats

- branch: Course information (for Polytechnic/Diploma programs)

#### 9. **diploma\_institutes\_enrollments.pq**

**Description**: Institute level enrollment data for Polytechnic / Diploma institutes.

**Key Fields**:

- sams\_code: Unique institute identifier

- academic\_year: Year of application

- module: Always “Diploma”

- institute\_name: Name of the institute

- category: Reservation category

- enrollments: Number of enrolled students

## Higher Secondary (HSS) Data Files (located in path (Cleaned/Higher Secondary))

### Student-Level Data

#### 10. **hss\_enrollment.pq**

**Description**: Student level enrollment data for higher secondary schools (HSS).

**Key Fields**:

- barcode: Unique student identifier

- academic\_year: Year of application (2018-2024)

- gender: Student gender

- social\_category: Caste category

- state, district: Location information

- module: Always “HSS”

#### 11. **hss\_applications.pq**

**Description**: Application level data from applications to higher secondary program seats.

**Key Fields**: - sams\_code: Unique institute identifier  
- barcode: Unique student identifier

- academic\_year: Year of application  
- num\_applications: Number of applications  
- option\_no: Student’s preference order for the institute (1 = first choice, 2 = second, etc.)  
- reported\_institute: Name of the institute they are reported to  
- stream: Chosen stream (Arts, Science, Commerce, etc.)

#### 12. **hss\_marks.pq**

**Description**: Student level data containing fields describing academic performance in the class 10 exams for students enrolled in higher secondary education

**Key Fields**: - barcode: Unique student identifier  
- academic\_year: Year of application  
- compartmental\_status: Whether the student has a compartment in any subject (Yes/No)  
- subject: Name of the subject  
- failed\_mark: Marks obtained in failed subject(s) - pass\_mark: Minimum passing marks

## Notes

### Privacy Protection

* All student national IDs (aadhar\_no) are encrypted using SHA-256 hashing

### Data Coverage

* **ITI**: 2017-2024 (8 years)
* **Diploma**: 2018-2024 (7 years)
* **HSS**: 2018-2024 (7 years)

Use the comprehensive README on the [GitHub repository](https://github.com/Data-Policy-and-Innovation-Centre/sams/tree/plus2-data) for more information.

## Technical Support

For questions about data structure, content, or analysis, please contact the DPIC team. Include the specific file name and your use case when requesting support.