

# Data Generation Documentation

## Overview

This document provides detailed information on the mock data generated for the **First Tern** session of the Vine Secondary School. The data represents a private secondary school in Delta State, Nigeria, with 600 students, 12 classes, 10 subjects per class, and 60 teachers. The data was generated to simulate real-world conditions in an African educational context, reflecting various aspects of student performance, teacher involvement, class structures, extracurricular activities, and parent information. The data is designed to support the development of a data warehouse for reporting, analysis, and decision-making.

## Data Generation Tools

To create this mock dataset, the following tools were used:

**Faker:** A Python library used to generate synthetic data such as names, birth dates, addresses, and other personal information.

**Python random module:** Utilized to create randomized distributions for student grades, attendance, extracurricular activities, and parental information.

**Pandas:** A Python library used to structure and manipulate the generated data, preparing it for loading into CSV files.

**Numpy:** Assisted with data manipulations and random selections.

**CSV Files:** The data was saved in CSV format, ready for loading into the data warehouse.

## School Structure

**Type:** Private secondary school located in Delta State, Nigeria.

**Student Population:** 600 students, split evenly between junior and senior secondary levels.

**Junior Secondary (JS1 to JS3):** 6 classes (JS1A, JS1B, JS2A, JS2B, JS3A, JS3B)

**Senior Secondary (SS1 to SS3):** 6 classes (SS1A, SS1B, SS2A, SS2B, SS3A, SS3B)

**Class Size:** 50 students per class arm, with a total of 12 class arms across both levels.

**Subjects:** Each class level has 10 subjects, with a mix of arts, sciences, and vocational subjects for junior and senior secondary students.

**Teachers:** A total of 60 teachers, with each teacher assigned to teach specific subjects across different class levels.

## Key Data Points

The data generated includes the following key aspects:

### 1. Students

Total: 600 students

Attributes: Student name, gender, age, date of birth, enrollment date (randomly set between February 7th and February 21st, 2024), and parent ID (some parents have more than one child in the school).

Class Assignment: Each student is assigned to a class arm based on their grade level.

### 2. Parents

Attributes: Name, occupation, education level, income bracket, and contact information.

Relationship: One parent may have multiple children (students) in the school.

### 3. Teachers

Total: 60 teachers

Attributes: Name, date of birth, subject expertise, qualification, years of experience, hire date, and salary.

Subject Allocation: Each teacher is assigned to specific subjects and classes.

#### **4. Classes**

Structure: 12 class arms, split between junior secondary (JS1 to JS3) and senior secondary (SS1 to SS3).

Attributes: Class level, class arm, and assigned subjects.

#### **5. Subjects**

Junior Secondary: Subjects such as Mathematics, English, Basic Sciences, and more.

Senior Secondary: Subjects such as Mathematics, Biology, Chemistry, Literature, and more.

Total: 10 subjects per class level.

#### **6. Grades**

Exams: Each student has grades for 10 subjects, with 15 marks for the first test, 15 marks for the second test, and 70 marks for the exam.

Attributes: First test, second test, exam score, total score, and final grade (A, B, C, D, E, or F).

#### **7. Attendance**

Period: Attendance data was generated for 3 months, starting from March 6th, 2024, with daily attendance marked from Monday to Friday.

Attributes: Student ID, attendance date, and status (Present or Absent).

#### **8. Extracurricular Activities**

Activities: Students participate in activities such as Drama, Music, Football, Dance, Basketball, and Volleyball.

Schedule: Each activity occurs every Friday, with students participating for 1, 2, or 3 hours

#### **9. Class-Subject Allocation**

Attributes: Class ID, subject ID, teacher ID, academic year, and periods per week (4 periods per week for each subject).

## **10. Survey Data**

Survey: A student survey was conducted at the end of the first term to understand various factors affecting their performance, such as study hours, access to resources, computer literacy, stress levels, exam anxiety, and motivation.

Anonymity: The survey was anonymous and included 600 responses.