**Project Blueprint: ZHI RISE Program Impact Analytics**.

**Project Title: Impact Analytics Dashboard for the RISE Initiative**

**The Scenario:** You are analyzing the performance of the **RISE program**, which targets Adolescent Girls and Young Women (AGYW) to reduce HIV risk. The program uses multiple interventions: **Economic Strengthening** (giving them financial skills/funds), **Education Support** (keeping them in school), and **Clinical Services** (PrEP, HIV testing).

**The Business Problem:** The program director needs to know: *"Which combination of interventions is most effective at keeping high-risk girls HIV-negative and in school?"* **The Data Analyst/Officer Task:**

1. **Data Quality Check (Python/SQL):** Identify missing beneficiary dates of birth, duplicate entries across districts, and mismatched intervention codes (Data Officers do a lot of this).
2. **Impact Analysis (Tableau):** Visualize the retention rate of girls in the program vs. their health outcomes.

---

**Proposed Synthetic Dataset**

I can generate a Python script to create a CSV file with the following columns (approx. 2,000 - 5,000 rows):

- **Beneficiary_ID:** Unique identifier.
- **District:** (Using ZHI locations: e.g., Gweru, Bulawayo, Mazowe).
- **Age_Group:** (10-14, 15-19, 20-24).
- **Intervention_Type:** (Economic Strengthening, Education Support, Sexual Violence Prevention, Clinical Only).
- **Enrollment_Date:** Date they joined the program.
- **HIV_Status_Baseline:** Status at start (Negative/Positive).
- **HIV_Status_Endline:** Status after 12 months.
- **School_Attendance_Rate:** % of classes attended (to measure "Education Support" impact).
- **Program_Status:** (Active, Graduated, Dropped Out).

**Why this works for the "Data Officer" role:**

- It uses **Health/NGO domain language** (Baseline/Endline, Interventions, Beneficiaries).
- It demonstrates **Impact Analytics** (correlating interventions with health outcomes).
- It allows you to show off **SQL** (querying specific districts) and **Tableau** (mapping the districts).

**Project Objective**

Analyze the effectiveness of the **RISE (Re-Ignite, Innovate, Sustain, and Empower)** program. We will determine which interventions (Education vs. Economic vs. Clinical) best improve retention rates and health outcomes for Adolescent Girls and Young Women (AGYW).

---

**Phase 1: Advanced Data Generation (Python)**

*Goal: Create a "Smart" Synthetic Dataset (approx. 5,000 records).* We will use Faker and NumPy with specific business logic to ensure realism.

- **Logic Rule 1 (The "Impact" Correlation):** Beneficiaries with **Multiple Interventions** (bundled services) will have higher "Retention Rates" than those with single interventions.
- **Logic Rule 2 (Geographic Realism):** We will use real ZHI districts (e.g., Gweru, Bulawayo, Bubi, Gokwe South) with varying population densities.
- **Logic Rule 3 (Time-Series Consistency):** "Endline" dates will always be logically after "Baseline" dates.

**The Datasets (3 Tables):**

1. **Beneficiaries_Table:** (ID, Age, District, Enrollment_Date, Vulnerability_Score).
2. **Interventions_Log:** (Beneficiary_ID, Intervention_Type, Date_Provided, Cost).
   - *Types:* School Fees Support, Financial Literacy Training, PrEP Initiation, SGBV (Sexual & Gender-Based Violence) Counseling.
3. **Impact_Outcomes:** (Beneficiary_ID, Baseline_HIV_Risk_Score, Endline_HIV_Risk_Score, School_Attendance_Baseline, School_Attendance_Endline, Status).

---

**Phase 2: SQL Data Warehousing & Transformation**

*Goal: Simulate the "Data Officer" role of managing data integrity.* We will load the CSVs into a SQL environment to perform:

1. **Data Cleaning:** Identify and flag "ghost beneficiaries" (IDs in the intervention log that don't exist in the master table).
2. **Feature Engineering:** Create a "bundled_service_flag" (Did they get 1 service or 3?).
3. **KPI Calculation:**
   - *Retention Rate:* % of beneficiaries active > 12 months.
   - *Impact Score:* (Endline School Attendance - Baseline School Attendance).

**Step 1: Load "Raw" Data to MySQL (The "EL" in ELT)**

We take the CSVs we just generated and load them into MySQL as-is.

- *Table 1:* raw_beneficiaries
- *Table 2:* raw_interventions

**Step 2: SQL Transformation (The "T" in ELT)**

We write SQL scripts to "clean" and "model" the data. We will write queries to:

- **Join** the beneficiaries to their interventions.
- **Aggregate:** Calculate "Total Spend per Beneficiary" (Summing up the costs).
- **Flag:** Create the "Dropout Flag" (1 or 0).
- *Result:* We create a final "Golden Table" or View called analytics_master_table.

**Step 3: Python EDA & Modeling**

We connect a Jupyter Notebook to MySQL.

- We read analytics_master_table.
- We run the Logistic Regression (predicting who will drop out).
- We save the predictions back to a new SQL table: predicted_dropout_risks.

**Step 4: Tableau Dashboarding**

We open Tableau and connect to MySQL.

- We join analytics_master_table with predicted_dropout_risks.
- We build the charts.

---

### Phase 3: Machine Learning (Predictive Risk Modeling)

*Goal: Add the "Impact Analytics" edge.* Since you have a Data Science background, we will add a lightweight **Logistic Regression** model.

- **The Problem:** "Can we predict which girls are likely to drop out of the program before it happens?"
- **The Output:** We will generate a **"Dropout Risk Score" (0-100)** for every active beneficiary. This allows the dashboard to show a "High Risk Alert" list.

---

### Phase 4: Advanced Tableau Visualization

*Goal: Move beyond basic bars/lines to "Decision Intelligence".* We will build **2 Interactive Dashboards**.

#### Dashboard 1: Strategic Impact Overview (The "Director's View")

- **Target Audience:** Program Director / Donors.
- **Key Question:** "Is the program working?"
- **Visuals:**
    - **KPI Banners:** Total Lives Touched, Retention Rate, % Risk Reduction.
    - **Map Visualization:** Zimbabwe District Map with a "Heatmap" of HIV Risk reduction (Darker Green = Better Impact).
    - **Sankey Diagram (or Flow Bar):** Showing the flow of Status (Enrolled -> Intervened -> Graduated/Dropped).

#### Dashboard 2: Operational "Risk & Action" Board (The "Field Officer's View")

- **Target Audience:** District Managers.
- **Key Question:** "Who do I need to visit today?"

- **Visuals:**
  - **Risk Quadrant Scatterplot:** *Cost of Intervention* (X-axis) vs. *Impact Score* (Yaxis). (Identifies high-cost/low-impact cases).
  - **"At-Risk" Table:** A filtered list of beneficiaries with a high "Dropout Risk Score" (from our ML model) so officers can intervene early.
  - **Parameter Filter:** "What-If" Analysis (e.g., "If we increase Education budget by 10%, how does retention change?").