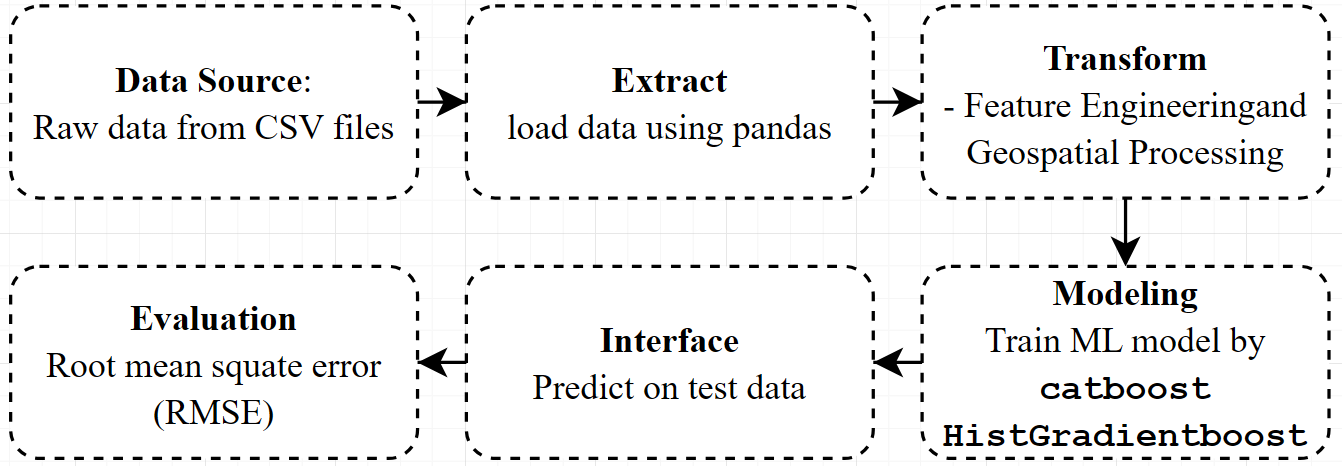
#### **International Women's Day Challenge**

## **Overview and Objectives**

The present project is oriented towards the development of predictive models utilizing geographic data is to construct a predictive model that accurately estimates the percentage of households per ward that objective female-headed and living below a specified income threshold. This will be achieved by using data points that can be collected through other means without an intensive household survey, such as the census.

## **Architecture Diagram**



## **ETL Process**

### **Extract**

The dataset is extracted from external CSV files hosted on GitHub. It includes training data, test data, sample submissions, and variable definitions.

### **Transform**

Data transformation involves preprocessing, handling missing values, and feature engineering. This includes geospatial transformations using H3 hexagons, distance calculations, and encoding categorical variables.

### **Load**

The transformed data is used directly for training machine learning models. It is temporarily stored in Pandas DataFrames during processing.

### **Data Modeling**

The model utilizes feature engineering techniques, including geospatial processing with H3 hexagons and distance calculations. The training process involves multiple algorithms, including **CatBoost**, and **HistGradientBoost**.

Model evaluation is performed using Root Mean Squared Error (RMSE).

### **Inference**

The trained model is used to make predictions on test data. The output is structured as two columns included the target

### **Run Time**

Total Notebook Execution Time is **221.68 second** which is around **3.69 minutes** to run. This includes the time taken for data loading, preprocessing, model training, and evaluation. The time may vary based on the system and environment. I run the notebook local

### **Performance Metrics**

Evaluation is based on Root Mean Squared Error (RMSE). KFold is used to ensure model reliability. The public score is **3.47** and the private is **3.50**