EHR Data Quality Auditor - Project Overview

About the Project

The EHR (Electronic Health Record) Data Quality Auditor is a Python-based tool that analyzes healthcare datasets in CSV format and identifies issues related to:

- Completeness - Are required fields like Age, DiagnosisCode filled?
- Consistency - Do fields follow correct formats (e.g., BloodPressure)?
- Correctness - Are values within valid clinical ranges?

The tool uses Pandas, Regex, and Streamlit for a lightweight, visual interface that lets users quickly inspect and validate datasets.

Main Purpose

To ensure that patient health records are clean, usable, and reliable before being used in:

- Medical decision-making
- Research and clinical trials
- Machine learning models
- Hospital dashboards

Bad data = bad decisions. This tool flags those issues early.

Benefits

- Fast detection of invalid and missing entries
- Improves trust in patient data
- Reduces time spent on manual validation
- Boosts accuracy of downstream medical research and AI pipelines
- Can be extended to include export features or integrate into EMR platforms

What It Checks

Check Type - Description - Example of Invalid

- Missing Values: Empty fields in Age, HeartRate (Age = NaN)
- Invalid Gender: Not in ['M', 'F', 'O'] (Gender = X)
- Out-of-Range Age: Age not between 0-120 (Age = 145)
- Heart Rate Errors: Non-numeric or outside 30-180 bpm (HeartRate = err)
- BP Format Check: Must be NNN/NNN (BP = invalid)
- Diagnosis Code: Must follow ICD-10 (like A01.1, B20) (Diagnosis = 123)

Sample Output

CSV Input Row Example:

PatientID: 1005

Age: 145

Gender: M

BloodPressure: 220/110

HeartRate: 105

DiagnosisCode: invalid

CLI Output Example:

Patient 1005 has issues: Unrealistic Age, Invalid Diagnosis Code

Streamlit Output includes:

- Data preview table
- Completeness percentage per column
- Problem rows with issue details
- Missing data matrix

Real-World Use Cases

- Hospitals: Validating EHR uploads before storage
- Researchers: Cleaning clinical trial datasets
- Data Scientists: Preparing medical data for ML training
- Auditors: Checking data standard compliance

Tech Stack

- Python 3.x
- Pandas - data processing
- Streamlit - interactive web dashboard
- Regex - pattern validation
- Faker - generates test datasets