
Phase 1 Report:

Quality Control

Team Name: *[Ahmed Ragheb(202301566), Ahmed Niazy(202300852) , Ibrahim Atef(202301058)]*
Dataset: Healthcare Dataset

Summary of Data Issues:

After analyzing the dataset during Phase 0, we identified the following issues that required cleaning:

- Missing values in columns like Age, Name, and Email
- Presence of duplicated rows
- Inconsistent text formatting (extra whitespace)
- Invalid data types for Age, Billing Amount, and Room Number
- No validation schema was applied to ensure data consistency

---

Cleaning Steps Taken:

To ensure the dataset is clean and ready for analysis, the following preprocessing steps were applied:

- Removed duplicate rows
- Trimmed whitespace from string/text columns
- Replaced placeholders like "N/A", "na", "" with proper NaN
- Converted the Age column to numeric and filled missing values with the median
- Filled missing string fields (e.g., Name) with "Unknown"
- Converted Billing Amount and Room Number to numeric types (as applicable)

Validation Schema:

We used the pandera library to define a schema to validate the dataset:

Column	Data Type	Checks
Name	String	Cannot be null
Age	Float	Must be >= 0
Gender	String	Must be either "Male" or "Female"
Billing Amount	Float	Must be >= 0
Room Number	Integer	Must be >= 0

The validation schema was used to enforce consistency and ensure the dataset was in a clean, usable state.

Output Files:

- phase1_cleaned_healthcare_dataset.csv: Cleaned dataset
- phase1_validation_script.py: Python code used for cleaning and validation

---

Conclusion:

The dataset has been successfully cleaned and validated according to best practices. All required Phase 1 deliverables are complete and submitted.
