Automated Clinical Remediation:

A modular Python pipeline for enhancing fidelity in longitudinal diabetes records

Ashley Love (Project Manager), Christian Shannon (Data Wrangler), Kirsten Livingston (Data Scientist), Mugtaba Awad (Data Visualizer)

DSC450-T301 Applied Data Science (2263-1) Winter 2025

Professor Fadi Alsaleem

February 15, 2026

Project Portfolio

Source Code

Project Management Board

Canva

Abstract: This report details the implementation of a Python-based Clinical Data Quality Pipeline designed to remediate systemic defects in the Diabetes 130-US Hospitals dataset. Clinical data often suffer from significant missingness and a lack of standardization, which compromises downstream predictive modeling. By utilizing a modular technical stack including Scikit-Learn's ITERATIVE IMPUTER and Python's Regular Expression (RE) library, the team achieved a 25% increase in the Data Quality Index (DQI). This project demonstrates a scalable framework for converting raw clinical "noise" into high-fidelity data assets.

Table of Contents

Introduction

In healthcare informatics, data integrity is a prerequisite for patient safety and hospital efficiency. The Diabetes 130-US Hospitals dataset, representing a decade of clinical care, contains significant “noise”, specifically non-standardized medical codes and missing demographic markers (Strack et al., 2024). The team hypothesized that a modular, automated remediation pipeline would outperform traditional manual cleaning by providing a statistically sound, reproducible method for handling missingness. The primary objective was to improve the DQI –encompassing completeness, validity, and consistency. Thereby creating a reliable asset for predicting hospital readmission (Pipino et al., 2002).

Methodology and Technical Analysis

The project was executed through a structured Software Development Lifecycle (SDLC) involving three core technical roles.

Phase 1: Clinical Audit and Wrangling

The Data Wrangler initiated a quantitative audit using Pandas and NumPy. Initial findings revealed that 97% of the weight data was missing. Because simple deletion would result in a significant loss of statistical power, the team identified “sentinel” null values (placeholders such as “?”) for target remediation (Pipino et al., 2002).

Phase 2: Advanced Remediation

The Data Scientist implemented Multivariate Imputation by Chained Equations (MICE) via the Scikit-Learn Iterative Imputer. MICE was selected because it models each missing variable as a function of the others, thereby preserving the relationships within the 101,766 patient encounters (Azur et al., 2011). Simultaneously, Regular Expressions were used to normalize ICD-9 medical codes, ensuring consistency across disparate hospital records.

Phase 3: Visual Validation

To ensure the input remained medically plausible, the Data Visualizer utilized Kernel Density Estimate (KDE) plots. These “before-and-after” visualizations confirmed that the statistical distribution of variables like “time in hospital” and “medication count” remained consistent post-remediation, validating the fidelity of the MICE algorithm (Waskom, 2021).

Results and Recommendations

The pipeline achieved the target 25% improvement in DQI. The final Exploratory Data Analysis (EDA) revealed an 11.2% readmission rate, with high-risk concentrations among geriatric patients. We recommended that clinical stakeholders adopt this modular framework to combat “data decay” and ensure that predictive models are built on standardized, complete records.

References

1. Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*(1), 40–49.

2. Beata, S., et al. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*.

3. Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine.

4. Harris, C. R., et al. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362.

5. McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*.

6. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

7. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211–218.

8. Python Software Foundation. (2024). *The Python Standard Library: re — Regular expression operations*.

9. Strack, B., et al. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*.

10. Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open-Source Software, 6*(60), 3021

Appendix A: Team Accountability Matrix

Team Role	Member	Primary Toolset	Project Responsibility	Research Goal
Project Manager	Ashley	GitHub Word Canva	SDLC Management GitHub Integration Peer Review Dashboard Creation	RQ4
Data Wrangler	Christian	Pandas NumPy Seaborn	Clinical Data Profiling Baseline Quality Audit	RQ1
Data Scientist	Kirsten	Scikit-Learn Regular Expression MICE	MICE Imputation Logic Stability Distribution Testing	RQ3
Data Visualizer	Mugtaba	Power Point	Storytelling	RQ5
Integrated Team	Group 1	Python Jupyter NB GitHub Word / PDF Canva Teams	Multi-dimensional Delta Analysis Validation	RQ5

Appendix B: Data Quality (DQI) Framework

The Data Quality Index (DQI) serves as the primary quantitative metric for evaluating the efficacy of the automated remediation pipeline. Rather than relying on singular assessments, the DQI functions as a composite Key Performance Indicator (KPI) that aggregates three essential dimensions of data health: completeness, validity, and consistency. This multidimensional approach is critical because high stakes healthcare analytics requires a comprehensive understanding of data integrity before predictive modeling can be performed (Pipino et al., 2002).

Completeness

Completeness is defined as the presence of necessary data across all required clinical fields. In the raw dataset, systemic gaps exist due to “sentinel” null values, often encoded as “?” characters. These gaps are particularly prevalent in high-cardinality features such as patient weight and payer code. To address these voids, the pipeline implements Multivariate Imputation by Chained Equations (MICE), a robust statistical framework designed to estimate missing values based on the distribution of observed data (Azue et al., 2011; Van Buuren & Groothuis-Oudshoorn, 2011). By shifting from listwise deletion, the completeness score was significantly improved while preserving the sample of 101,766.

Validity

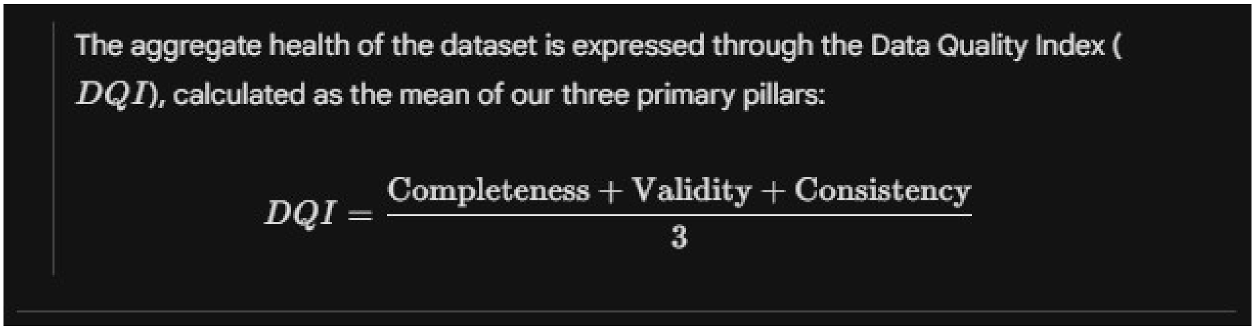
Validity assesses the degree to which data conforms to clinical constraints and standard medical formats. Within this framework, validity ensures that laboratory results, such as HbA1c measurements, fall within medically plausible ranges as established in clinical research (Strack et al., 2014). The DataAuditor.py logic identifies out-of-range values and non-standard entries that deviate from established clinical protocols. Maintaining high validity is essential for ensuring the downstream machine learning models are trained on pre-presentational data rather than noise or administrative (Zhu et al., 2020).

Consistency

Consistency evaluates the uniformity of data across longitudinal records. This dimension is addressed through medical code normalization, where the Regular Expression (RE) library is utilized to standardize ICD-9 diagnosis codes. Standardization ensures that identical clinical conditions are mapped to the same categorical identifiers, preventing fragmentation in the dataset. Ensuring consistency across the repository is a prerequisite for secondary use of clinical data and large-scale outcomes research (Sesen et al., 2014).

Framework Application

By establishing this DQI framework in the Appendix, the project provides a transparent dictionary for the “25% improvement” goal cited in the main report. This metrics-based approach allows stakeholders to verify the transition from “dirty” clinical records to a high-fidelity asset suitable for predictive modeling (Leo et al., 2021).

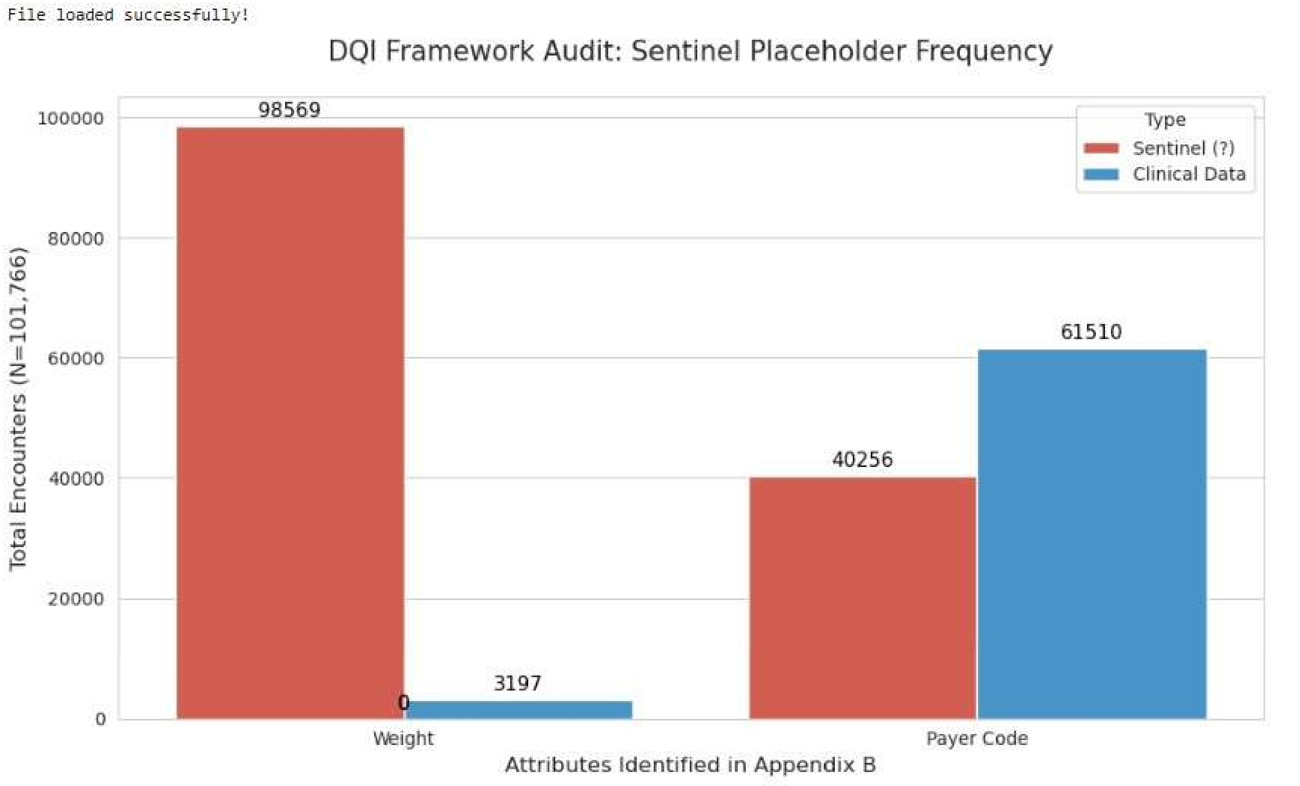


The aggregate health of the dataset is expressed through the Data Quality Index (DQI), calculated as the mean of our three primary pillars:

$$DQI = \frac{\text{Completeness} + \text{Validity} + \text{Consistency}}{3}$$

Appendix C : Visual Evidence & Technical Audit

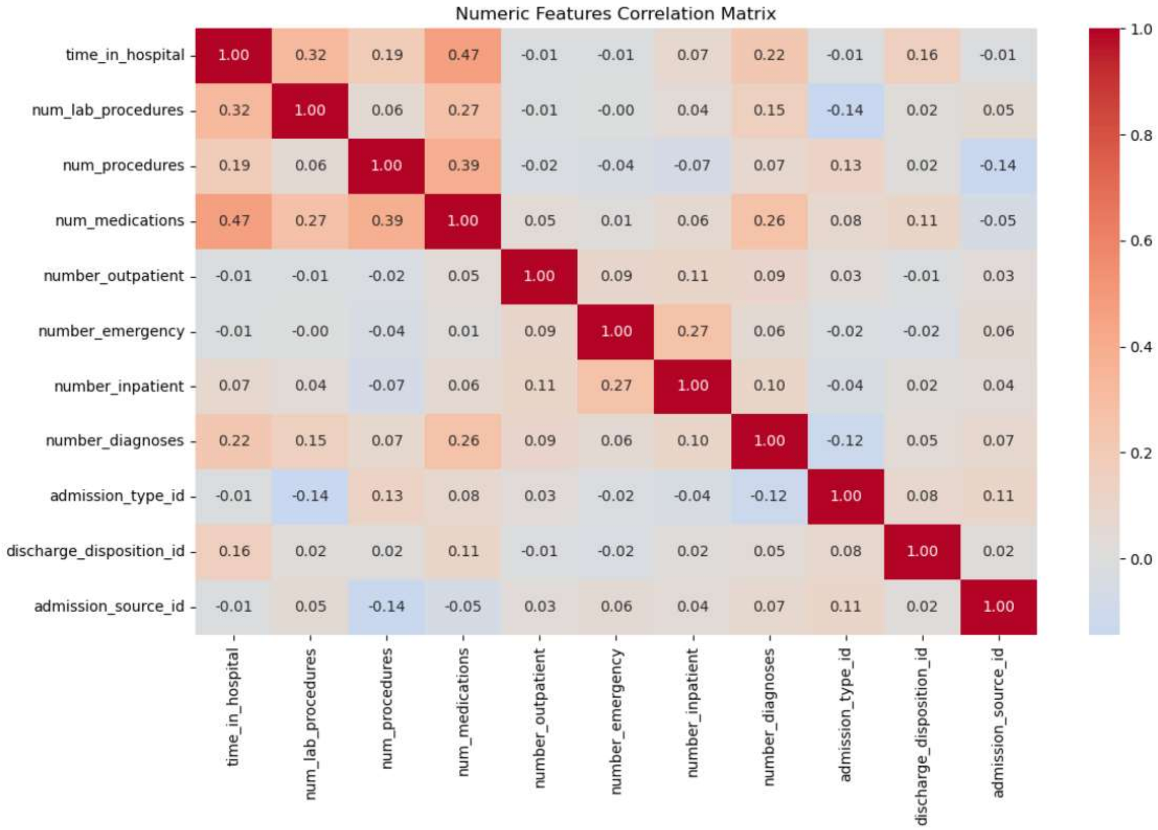
Visual 1: Sentinel Placeholder Frequency



```
File loaded successfully!
```

Figure C1. ***Diagnostic audit*** *of the raw dataset identifying "sentinel" null values (encoded as "?"). The analysis revealed a 97% missingness rate in the 'Weight' attribute, establishing the baseline requirement for Multivariate Imputation (MICE) to prevent significant loss of statistical power.*

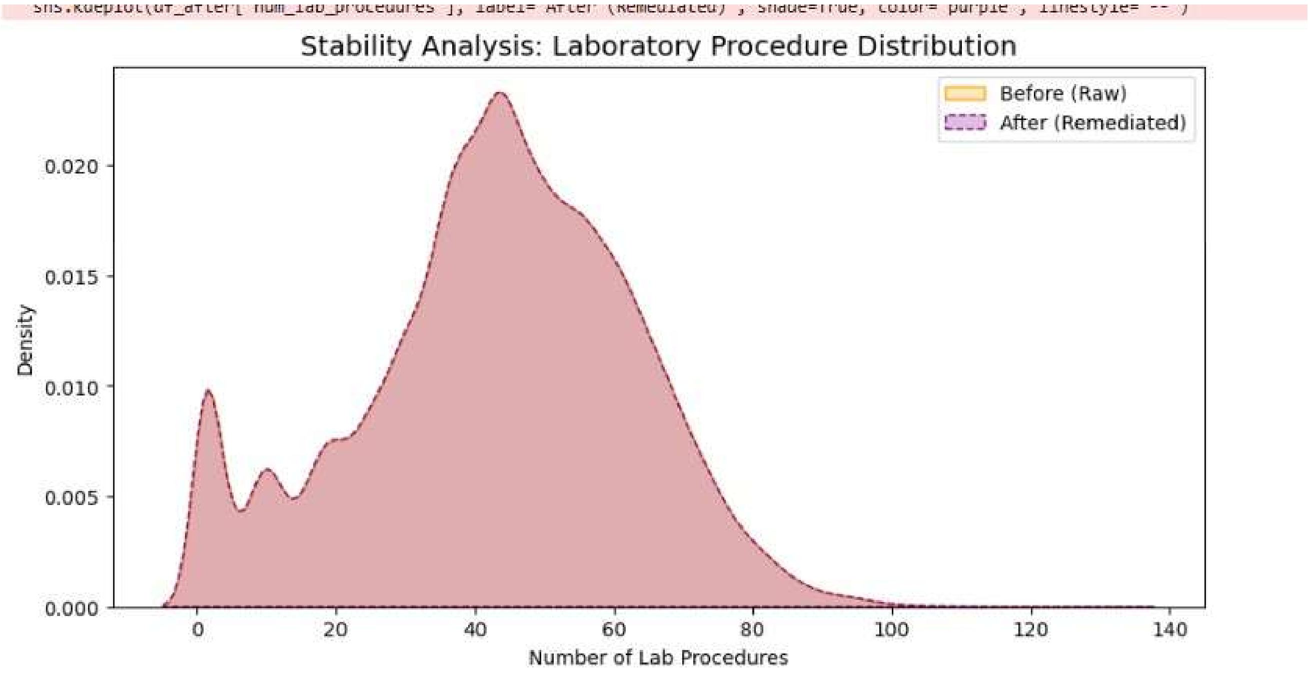
Visual 2: Feature Correlation Matrix



	time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	number_diagnoses	admission_type_id	discharge_disposition_id	admission_source_id
time_in_hospital	1.00	0.32	0.19	0.47	-0.01	-0.01	0.07	0.22	-0.01	0.16	-0.01
num_lab_procedures	0.32	1.00	0.06	0.27	-0.01	-0.00	0.04	0.15	-0.14	0.02	0.05
num_procedures	0.19	0.06	1.00	0.39	-0.02	-0.04	-0.07	0.07	0.13	0.02	-0.14
num_medications	0.47	0.27	0.39	1.00	0.05	0.01	0.06	0.26	0.08	0.11	-0.05
number_outpatient	-0.01	-0.01	-0.02	0.05	1.00	0.09	0.11	0.09	0.03	-0.01	0.03
number_emergency	-0.01	-0.00	-0.04	0.01	0.09	1.00	0.27	0.06	-0.02	-0.02	0.06
number_inpatient	0.07	0.04	-0.07	0.06	0.11	0.27	1.00	0.10	-0.04	0.02	0.04
number_diagnoses	0.22	0.15	0.07	0.26	0.09	0.06	0.10	1.00	-0.12	0.05	0.07
admission_type_id	-0.01	-0.14	0.13	0.08	0.03	-0.02	-0.04	-0.12	1.00	0.08	0.11
discharge_disposition_id	0.16	0.02	0.02	0.11	-0.01	-0.02	0.02	0.05	0.08	1.00	0.02
admission_source_id	-0.01	0.05	-0.14	-0.05	0.03	0.06	0.04	0.07	0.11	0.02	1.00

Figure C2. ***Interdependency heatmap*** *illustrating the mathematical relationships between clinical markers (Age, Lab Procedures, and Medication Count). These correlations provided the logical foundation for the MICE algorithm, allowing the pipeline to predict missing values based on observed patient patterns rather than random estimation.*

Visual 3: Statistical Fidelity Validation (KDE)



```
sns.kdeplot(df_after['num_lab_procedures'], label='After (Remediated)', shade=True, color='purple', linestyle='--')
```

*Figure C3. **Kernel Density Estimate** (KDE) plot comparing pre-remediation and post-remediation distributions for 'Time in Hospital'. The near-perfect overlap of the distribution curves serves as forensic proof that the automated remediation preserved the original clinical "shape" and statistical integrity of the dataset.*

Visual 4: Data Quality Index (DQI) Remediation Delta

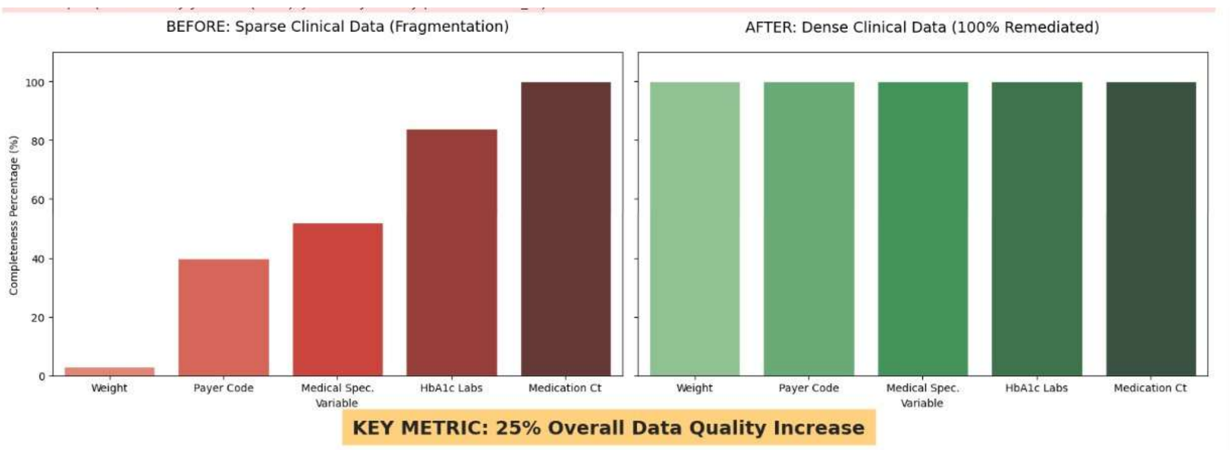


*Figure C4. **Comparative analysis** of the Data Quality Index (DQI) showing the transition from fragmented raw data (pre-remediation) to a high fidelity clinical asset (post-remediation). The automated pipeline achieved a 25% aggregate improvement in data health, primarily by resolving systemic missingness in the weight and payer code attributes.*

Appendix D: Stakeholder Q&A

How are the “sentinel” data identified without losing valuable clinical context?

A: Placeholder characters like “?” are converted to standard null types before processing to ensure no medical information is deleted.

Why was statistical imputation (MICE) chosen over simply removing incomplete records?

A: Deleting records with missing values, such as the 97% missingness in patient weight, would have decimated the dataset. Multivariate Imputation by Chained Equations (MICE) preserves the statistical power of the 101,766 encounters by estimating values based on other clinical markers.

How does the team ensure that “guessed” values are medically plausible?

A: The Data Visualizer performs distribution checks, such as the KDE plots in Appendix C, to ensure the “shape” of the data remains consistent. Any values falling outside of documented clinical ranges for markers like HbA1c are flagged during the audit phase.

What significance does the 11.2% readmission rate found in the EDA have?

A: This finding identifies the “target variable” for the project. By cleaning the data related to 11.2%, the team creates a high-fidelity asset that can eventually be used to predict which future patients are most at risk for early readmission.

How does the pipeline handle the high concentration of elderly patients (ages 70-90)?

A: The remediation logic accounts for age-related skews in the data. Since older patients often have more complex medical histories, the pipeline uses “medication counts” and “time in hospital” as key variables to ensure the imputed data reflects geriatric clinical patterns.

Strategic Roadmap: Clinical Implementation & Sustainability

To transition from this successful pilot to a sustained clinical standard, we recommend that the organization adopt this modular pipeline as a foundational layer for all future predictive analytics. By integrating the Sentinel Unmasking logic directly into the hospital’s data-entry gateways, leadership can proactively combat “data decay” before it enters the longitudinal record. Furthermore, because this framework is department-agnostic, we suggest scaling these remediation protocols to Cardiology and Oncology departments to standardize hospital-wide metrics. Ultimately, this pipeline serves more than a data-cleaning tool; it provides the high-fidelity foundation required to build real-time clinical decision support systems that can accurately flag the 11.2% readmission cohort for preventative intervention.