Prepared By:

Muzakkir Ansari

B.Sc Statistics, 2024

Mumbai, India

muzakkiransari001@gmail.com

SAS Data Cleaning & Analysis Project Report

1. Project Title

- Patient & Encounters Data Cleaning, Integration, and Analysis using SAS

2. Objective

- To demonstrate SAS programming skills in data cleaning, transformation, dataset integration, derivations, and statistical analysis, following industry best practices in clinical data and healthcare analytics.

3. Raw Datasets

- patients_raw.csv: Contains patient demographics.
- encounters_raw.csv: Contains patient visit information.

4. Data Issues Identified

- Patient IDs: inconsistent (P001, p-002, 003).
- DOB: multiple formats (12/03/1985, 1988-07-05, 15-Aug-1992), invalid dates (31/02/1987).
- Sex: mixed (M, male, FEMALE, blanks).
- Phone: varied formats, symbols, missing values.
- City: inconsistent case, abbreviations (mumbai, MUM).
- Department: abbreviations (ENT, Ortho), inconsistent names.
- Charges: stored as character with symbols.
- Status: inconsistent (Y, Yes, Completed).

5. Data Cleaning & Transformation Steps

Patients dataset:

- Standardized PatientsID → format P###.
- Converted DOB → SAS DATE9. format, corrected invalid values.
- Standardized Sex → M, F, U (Unknown).
- Cleaned Phone → extracted last 10 digits; missing values flagged.
- Standardized City → proper case, mapped abbreviations.

Encounters dataset:

- Standardized PatientsID to match patients dataset.
- Converted VisitDate into SAS date format.
- Cleaned Department names → standardized categories.
- Converted Charges → numeric with 2 decimals.
- Cleaned Status → standardized (Yes, No, Completed, Cancelled).

6. Dataset Integration

- Merged patients_final and encounters_final by PatientsID.
- Derived Age = intck("year", DateOfBirth, VisitDate).
- Created PaidFlag → 1=Paid, 0=Unpaid, .=Missing.

7. Analysis Performed

- Frequency of Department × Sex.
- Descriptive statistics of Charges by Department (Mean, Std Dev, Min, Max).
- Distribution of PaidFlag (Paid vs Not Paid).

8. Outputs Generated

- patients_final.csv → cleaned patient-level data.
- encounters_final.csv → cleaned visit-level data.
- merge_visits.csv → merged dataset with derived variables.
- Summary frequency and means tables.

9. Conclusion

- This project demonstrates the complete SAS data pipeline:
- Importing raw CSV files
- Cleaning and standardizing variables
- Merging datasets
- Deriving new clinical and operational metrics
- Performing descriptive analysis
- Exporting final datasets

The workflow follows industry practices of traceability, minimal overwriting, use of flags, and transparent derivation.