Data Intake Report

Name: Persistency of Drug – Healthcare Project
Report date: 19/08/2025
Internship Batch: LISUM47
Version: 1.0
Data intake by: Keval Jaysukhbhai Savaliya
Data intake reviewer:
Data storage location:
https://drive.google.com/file/d/1P_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/view

Tabular data details:

File Name	Healthcare dataset
Total number of observations	3424
Total number of files	-
Total number of features	69
Base format of the file	.xlsx
Size of the data	898 KB

Objective:

This report documents the quality checks and structure of the Healthcare dataset provided for analyzing **persistency of a drug**. The dataset encompasses patient demographics, provider attributes, clinical factors, comorbidities, adherence, and risk factors to support the development of a classification model that predicts patient persistence.

Proposed Approach:

1. **Data Cleaning:**
    a. **Null Value Check:**
        i. **Approach for check:**
            - Used **.isnull().sum()** in Pandas to identify any row has null values or not.
            - If duplicates were found, we would drop them using **.dropna()**

    b. **Duplicate Value Check:**
        i. **Approach for check:**
            - Used **.duplicated().sum()** in Pandas to identify any full-row or key-based duplicates.
            - If duplicates were found, we would drop them using **.drop_duplicates()**

2. **Column-Wise Format Fixes:**
    a. Standardize categorical variables (e.g., Race, Gender, Region).

b. Convert Age buckets into numerical or ordinal encoding.

3. **Assumptions:**
   - Patient ID (Ptid) is unique and primary key.
   - Persistency_Flag is the dependent variable (binary classification).
   - Clinical metrics such as T-Score and Risk Segments are consistent across patients.

4. **Outlier detection**
   - Outliers are expected in continuous variables, such as Age, T-score, and Dexa Scan Frequency.

5. **Correlation Observations**

   Persistency likely correlated with Age, Adherence, Risk Segments, and Comorbidity factors.