Data Cleansing and Transformation Report

Team Members: Apurwa Bhausaheb Sontakke, Vedant Wagh

Group Name - HealthData Innovators

Name – Apurwa Bhausaheb Sontakke

Email – sontakke.ap@northeastern.edu

Country - USA

College– Northeastern University

Specialization – Data Science

Github Link - https://github.com/apurwasontakke/Week-9-Data-Science-Healthcare-Project-Data-Cleansing-and-Transformation

Problem Description

Pharmaceutical companies face a significant challenge in understanding why patients continue or discontinue their prescribed medications. To address this, ABC Pharma has sought the help of an analytics company to automate the identification process of factors influencing drug persistency. The aim is to develop a classification model that predicts whether a patient will persist with a prescribed drug (Persistency_Flag).

Handling Missing Values

Upon examining the dataset, it was determined that there were no missing values present. Each column had the full 3424 entries, indicating that the dataset was complete with no gaps in the data.

Verification of Missing Values

A check was performed to confirm the presence of missing values. The dataset was found to have all its values intact, and therefore, no imputation for missing values was necessary.

Summary of Missing Values

Initial Inspection: An initial inspection of the dataset revealed that all columns had non-null entries.

Verification: A thorough verification confirmed that there were no missing values in the dataset.

Since there were no missing values, the focus of the data cleansing process was primarily on handling outliers to ensure the data was prepared for further analysis.

Outlier Detection and Handling

Two methods were employed to detect and handle outliers in the dataset:

Z-score Method: This method identifies outliers based on the standard deviations from the mean. Values with Z-scores above a certain threshold (typically 3) are considered outliers. These outliers were then replaced with the mean value of the column.

IQR Method: The Interquartile Range (IQR) method identifies outliers as values that fall below the first quartile (Q1) minus 1.5 times the IQR or above the third quartile (Q3) plus 1.5 times the IQR. These outliers were replaced with the median value of the column.

Verification of Outliers

After applying the outlier handling methods, the remaining outliers were counted to evaluate the effectiveness of each method. The results showed:

Z-score Method: 83 remaining outliers in the Dexa_Freq_During_Rx column.

IQR Method: 200 remaining outliers in the Dexa_Freq_During_Rx column.