**Data Cleansing and Transformation Report**

**Team Members: Apurwa Bhausaheb Sontakke, Vedant Wagh**

**Group Name - HealthData Innovators**

**Name – Vedant Wagh**

**Email – wagh.v@northeastern.edu**

**Country - USA**

**College– Northeastern University**

**Specialization – Data Science**

**Github Link -**

**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:

**Isolation Forest**

Isolation Forest is an ensemble learning technique specifically designed for anomaly detection. The basic idea behind Isolation Forest is to isolate observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. This process is repeated recursively to create a tree structure. Observations that require fewer splits to be isolated are considered anomalies because they deviate significantly from the norm**.**

**Key Points:**

Isolation: Anomalies are isolated faster than normal points because they are fewer and different.

Random Splits: The algorithm creates random splits, making it less sensitive to specific distributions.

Efficient: It works well with large datasets.

**Local Outlier Factor (LOF)**

Local Outlier Factor (LOF) is a density-based anomaly detection algorithm. It measures the local deviation of a data point with respect to its neighbors. The idea is to compare the local density of a point to the local densities of its neighbors. Points that have a substantially lower density than their neighbors are considered anomalies.

**Key Points:**

Local Density: It uses the concept of local density to identify anomalies.

Neighbors: It considers a fixed number of nearest neighbors to calculate the density.

Relative Density: Anomalies are identified based on how much lower the density of a point is compared to its neighbors. Verification of Outliers