Problem Description

The objective of this analysis was to clean and transform a healthcare dataset to prepare it for further analysis and predictive modeling. The dataset contains various attributes related to patient demographics, medical history, risk factors, and treatment details. Key tasks included handling missing values, identifying and addressing outliers, transforming skewed data, converting data types, encoding categorical variables, and performing correlation analysis

Key Findings from Exploratory Data Analysis (EDA)

1. **Data Overview:**
   - The dataset contains 3424 entries and 69 columns, with a mix of numerical and categorical variables.
   - Important columns include patient ID (Ptid), target variable (Persistency_Flag), and various risk factors.
2. **Missing Values Analysis:**
   - There were no explicit missing values detected in the dataset. All columns had complete entries.
3. **Outlier Detection and Handling:**
   - Outliers in numerical columns (Dexa_Freq_During_Rx and Count_Of_Risks) were detected using the Z-score method.
   - Data points with Z-scores beyond a threshold of 3 were removed, reducing the dataset size to 3344 entries.
4. **Skewness in Data:**
   - Skewed data in numerical columns (Dexa_Freq_During_Rx and Count_Of_Risks) were transformed using log transformation to reduce skewness and normalize the distribution.
5. **Data Type Conversion:**
   - Categorical columns were converted to the category data type to ensure efficient memory usage and proper handling during analysis.
6. **Encoding Categorical Variables:**
   - One-hot encoding was applied to categorical variables to convert them into a binary matrix, making them suitable for machine learning algorithms.
7. **Correlation Analysis:**
   - The correlation matrix was calculated to identify relationships between the target variable and the feature variables.
   - Key features positively correlated with Persistency_Flag include Count_Of_Risks, Risk_Chronic_Liver_Disease_Yes, and Risk_Poor_Health_Frailty_Yes.

    - Key features negatively correlated with Persistency_Flag include Dexa_Freq_During_Rx_Log, Risk_Vitamin_D_Insufficiency_Yes, and Risk_Low_Calcium_Intake_Yes.

8. **Univariate and Bivariate Analysis:**
    - Histograms and count plots were used to visualize the distribution of numerical and categorical features.
    - Box plots and count plots by Persistency_Flag were used to explore relationships between features and the target variable.

Final Recommendation

The dataset has been successfully cleaned and transformed, making it ready for further analysis or modeling. Key steps included:

- Handling outliers using the Z-score method.
- Transforming skewed data with log transformation.
- Converting data types.
- Encoding categorical variables.
- Performing correlation analysis to identify significant features.

Based on the analysis:

- Features like Count_Of_Risks, Risk_Chronic_Liver_Disease_Yes, and Risk_Poor_Health_Frailty_Yes are highly positively correlated with the target variable and could be critical for predictive modeling.
- Features like Dexa_Freq_During_Rx_Log, Risk_Vitamin_D_Insufficiency_Yes, and Risk_Low_Calcium_Intake_Yes are negatively correlated with the target variable and also provide valuable information.