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.

Data Understanding

The dataset consists of 3424 entries and 69 columns, with a mix of numerical and categorical variables. Some of the key columns include patient ID (**Ptid**), target variable (**Persistency_Flag**), and various risk factors.

Data Cleaning and Transformation Process

Step 1: Import Libraries

We started by importing the necessary libraries for data analysis and visualization.

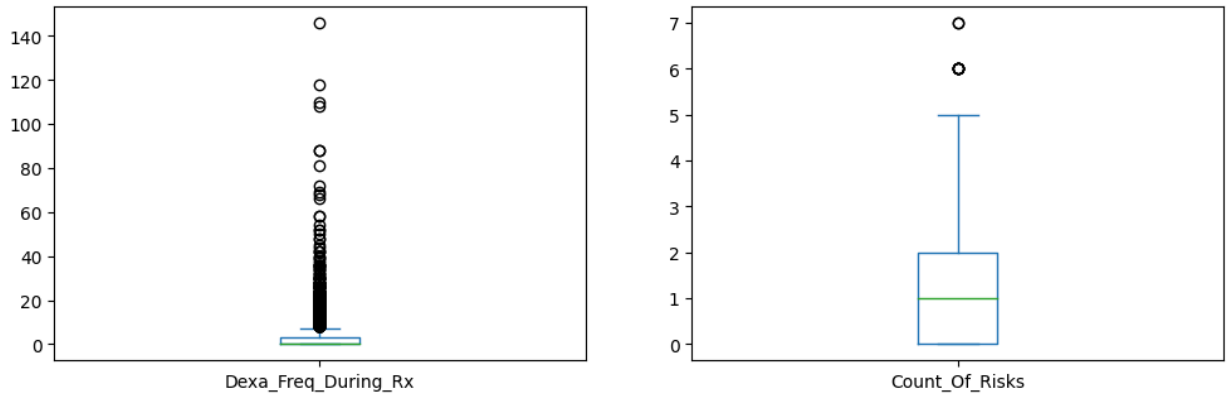
Step 2: Load Dataset

The dataset was loaded from an Excel file, and the data types of the categorical columns were converted to the **category** data type for efficient memory usage and proper handling.

Step 3: Outlier Detection and Handling

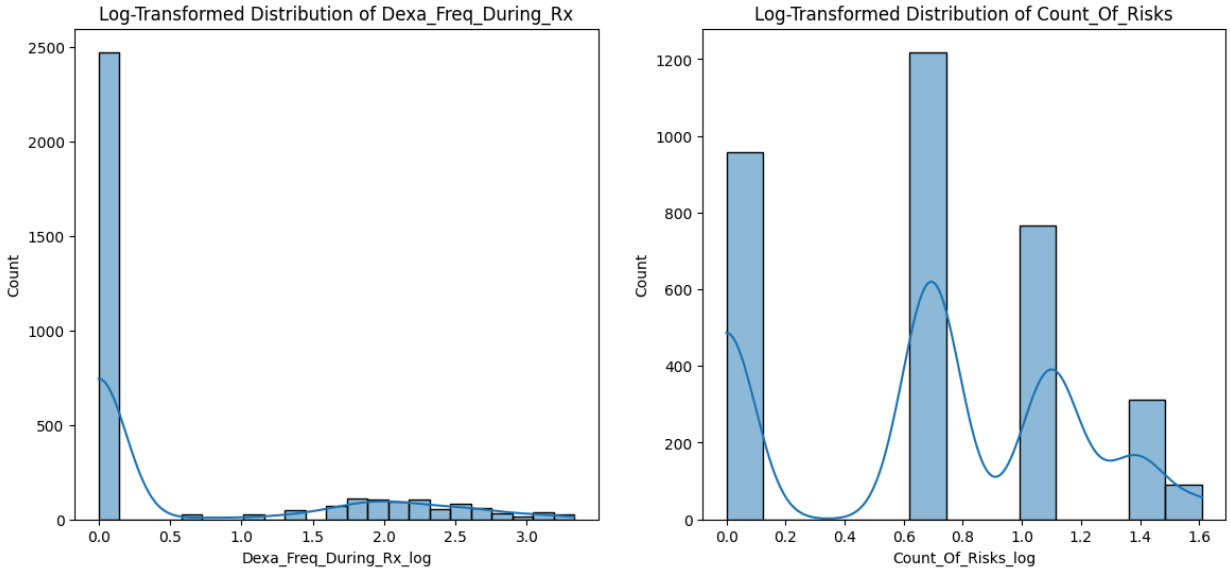
Outliers in numerical columns (**Dexa_Freq_During_Rx** and **Count_Of_Risks**) were detected using the Z-score method. Outliers are data points that are significantly different from other observations and can affect the results of the analysis.

This step reduced the dataset size to 3344 entries by removing the outliers.



Step 4: Transform Skewed 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.



Step 5: Convert Data Types

The data types of categorical columns were converted to the category data type to ensure efficient memory usage and proper handling during analysis.

Step 6: Separate Target Variable and Feature Variables

The target variable (Persistency_Flag) was separated from the feature variables. The Ptid column, which is an identifier and not useful for analysis, was dropped.

Step 7: Encoding Categorical Variables

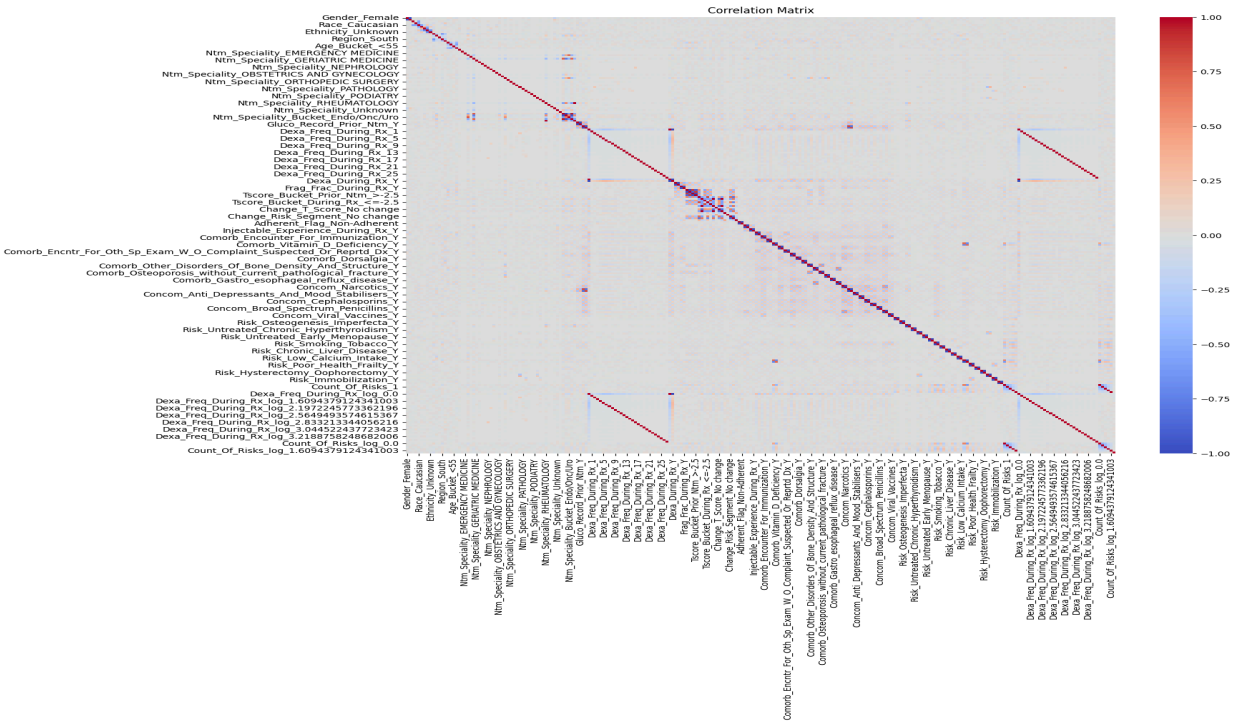
A function was defined to apply one-hot encoding to the categorical variables. One-hot encoding converts categorical variables into a binary matrix, making them suitable for machine learning algorithms.

Step 8: Combine Encoded Features and Target Variable

The target variable was converted to a numerical format, and the encoded feature variables were combined with the target variable into a single DataFrame for correlation analysis.

Step 9: Correlation Analysis

The correlation matrix was calculated to identify relationships between the target variable and the feature variables. The correlation matrix helps in understanding which features are most associated with the target variable.



Conclusion

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, and performing correlation analysis to identify significant features.

The correlation analysis revealed important features positively and negatively associated with the target variable, Persistency_Flag, which can be useful for building predictive models.