EDA 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-EDA

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).

Exploratory Data Analysis (EDA)

Load the Dataset

The dataset was successfully loaded and contains 3424 entries and 69 columns. The data includes various demographic, medical, and risk-related attributes.

Data Cleaning

Missing Values

An examination of the dataset revealed 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.

Descriptive Statistics

Numerical Columns

Descriptive statistics for numerical columns were calculated to understand the central tendency, dispersion, and shape of the dataset's distribution. Key statistics such as mean, standard deviation, minimum, maximum, and quartiles were obtained.

Key Findings:

Dexa_Freq_During_Rx: Mean = 3.02, Std = 8.14, Max = 146.

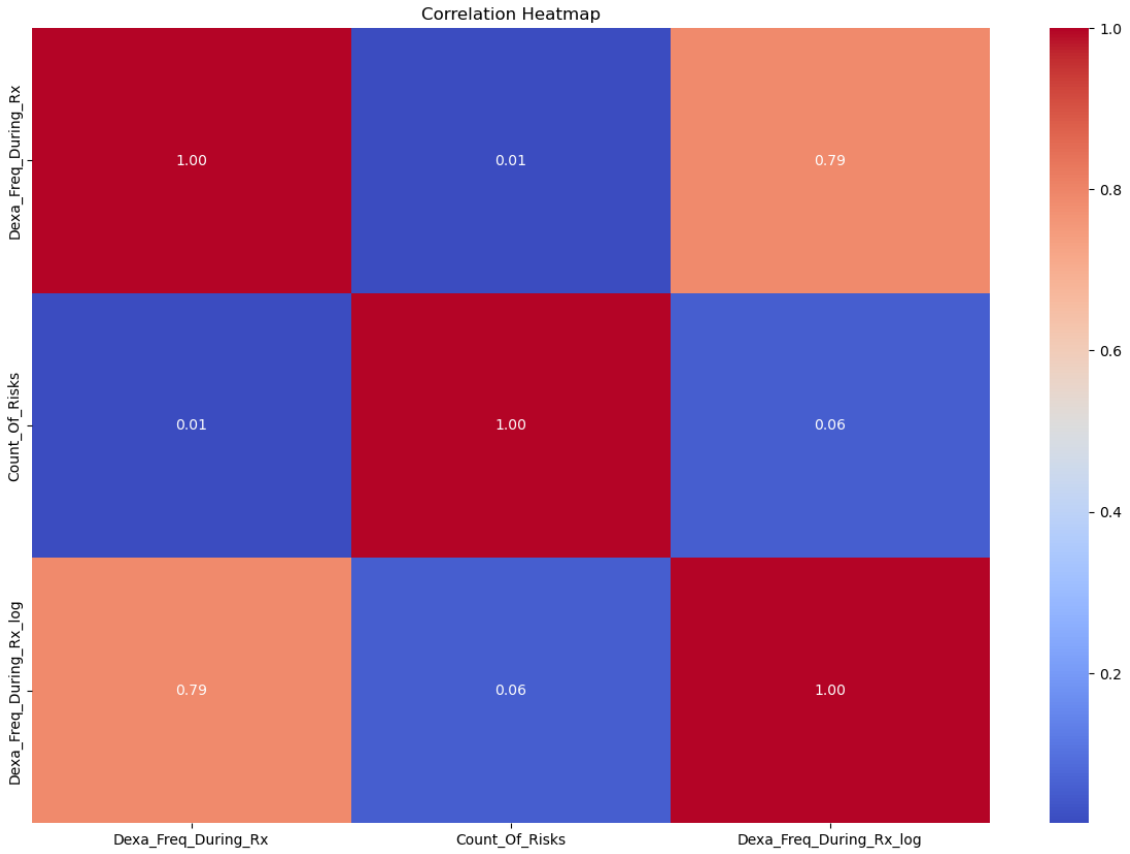
Count_Of_Risks: Mean = 1.24, Std = 1.09, Max = 7.

Categorical Columns

Frequency counts for categorical columns were analyzed to understand the distribution of categorical data. This helped in identifying the most common categories and potential imbalances in the data.

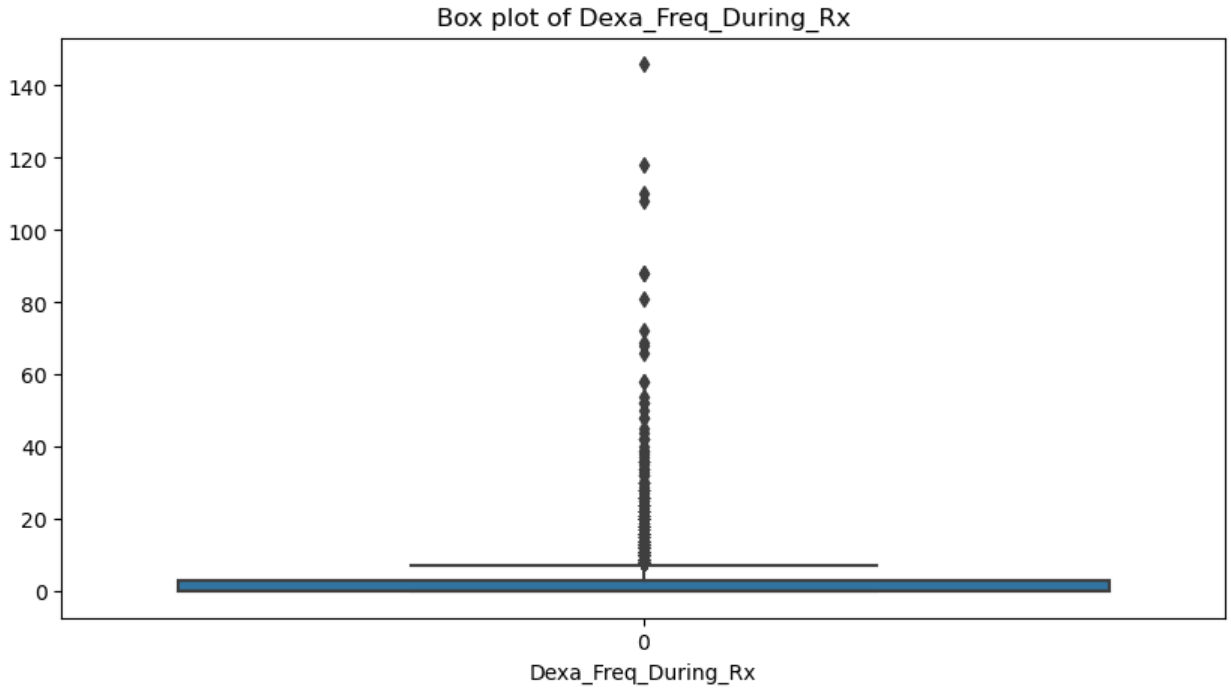
Data Visualization

The correlation heatmap displays the strength and direction of relationships between numerical variables in the dataset. Each cell in the heatmap shows the correlation coefficient between two variables, with values ranging from -1 to 1.



Key Observations:

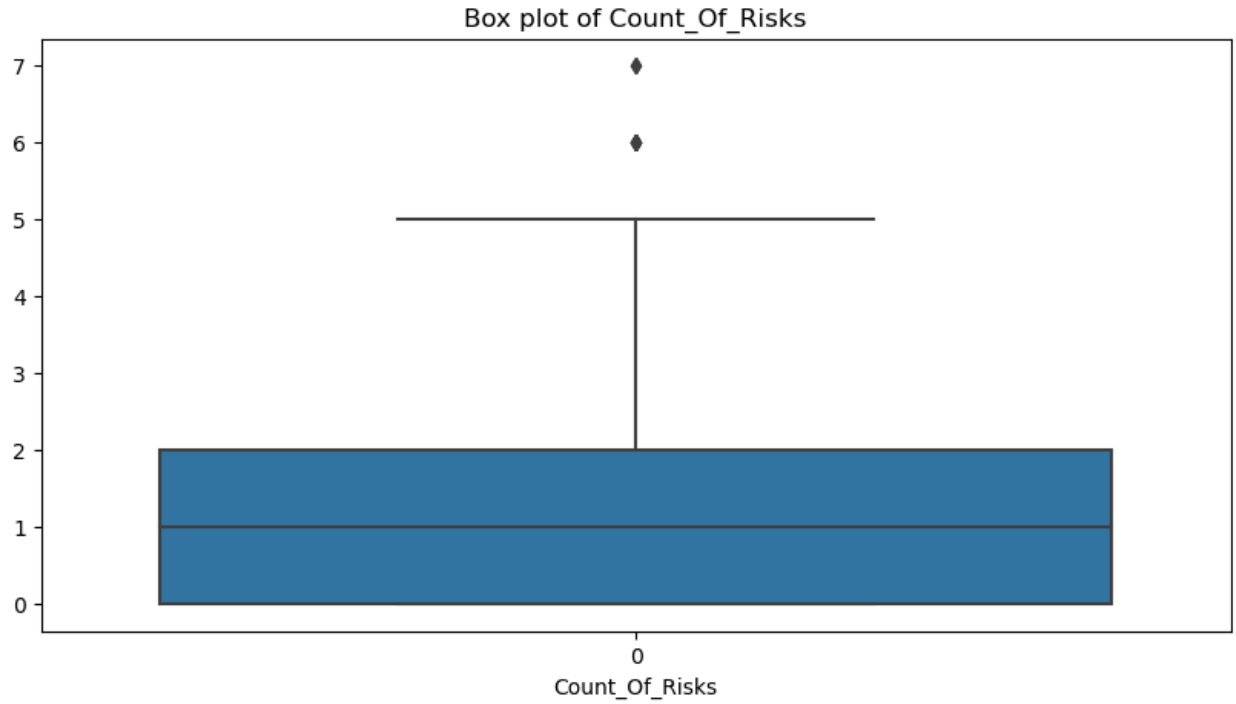
- **Dexa_Freq_During_Rx** and **Dexa_Freq_During_Rx_log** have a high positive correlation (0.79), indicating that the log transformation retains the relative order of values while reducing skewness.
- **Dexa_Freq_During_Rx** and **Count_Of_Risks** have a very weak positive correlation (0.01), indicating little to no linear relationship.
- **Count_Of_Risks** and **Dexa_Freq_During_Rx_log** also have a very weak positive correlation (0.06), similar to the original **Dexa_Freq_During_Rx** variable.



This box plot visualizes the distribution of the **Dexa_Freq_During_Rx** variable, highlighting the median, quartiles, and potential outliers.

Key Observations:

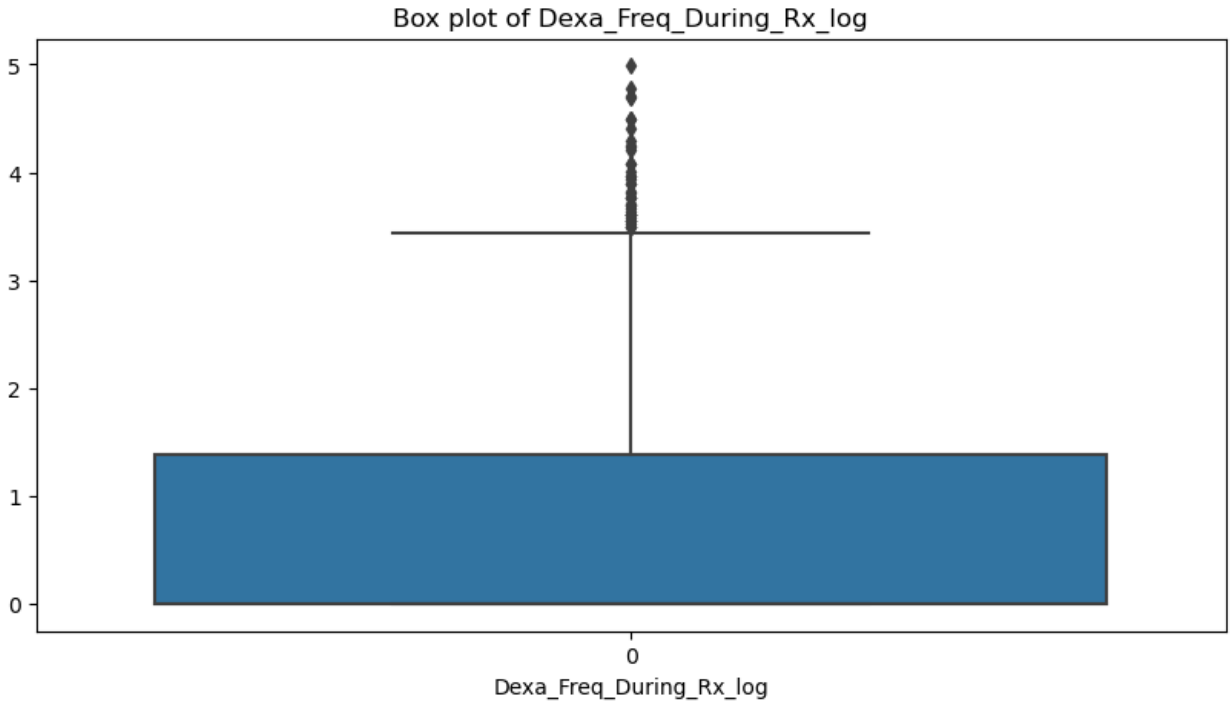
- The majority of the data points for **Dexa_Freq_During_Rx** are concentrated near the lower end of the scale.
- There are numerous outliers extending far beyond the upper quartile, indicating a right-skewed distribution with extreme values.



This box plot visualizes the distribution of the **Count_Of_Risks** variable, highlighting the median, quartiles, and potential outliers.

Key Observations:

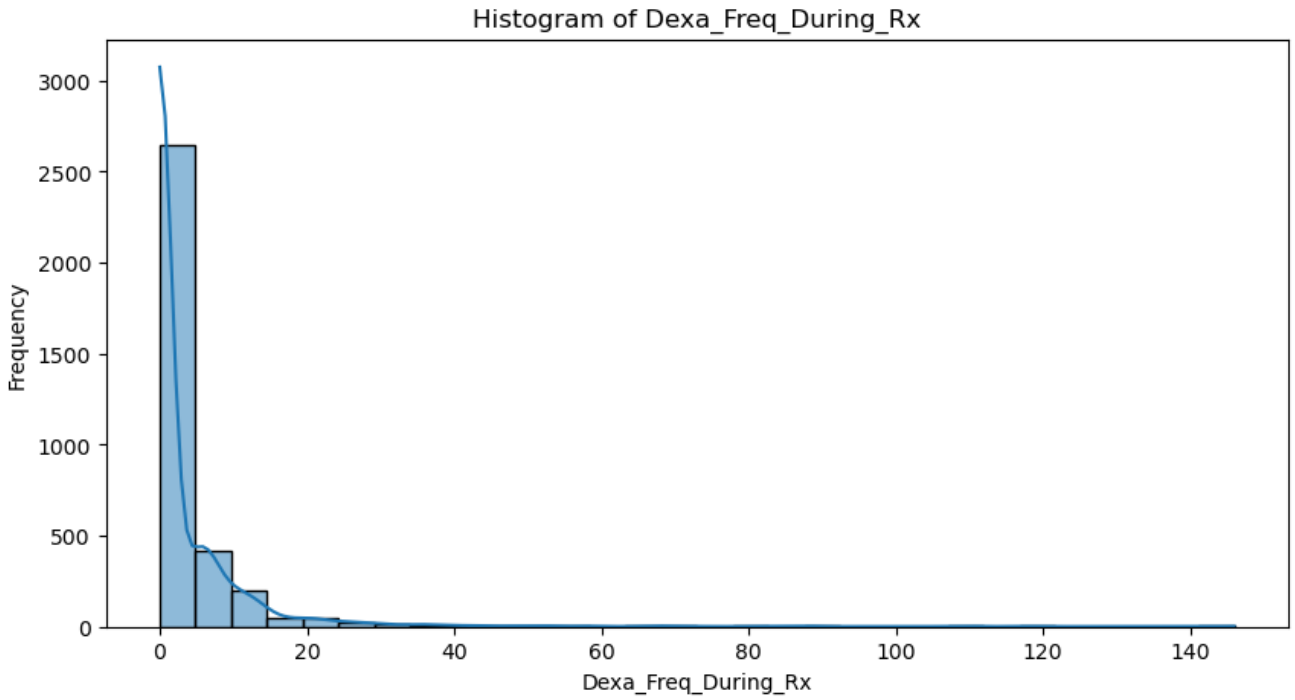
- The **Count_Of_Risks** variable has a more balanced distribution compared to **Dexa_Freq_During_Rx**.
- The median value is 1, and the interquartile range (IQR) is between 0 and 2.
- There are a few outliers beyond the upper whisker, indicating some higher risk counts that deviate from the typical range.



Description: This box plot visualizes the distribution of the log-transformed **Dexa_Freq_During_Rx** variable, highlighting the median, quartiles, and potential outliers.

Key Observations:

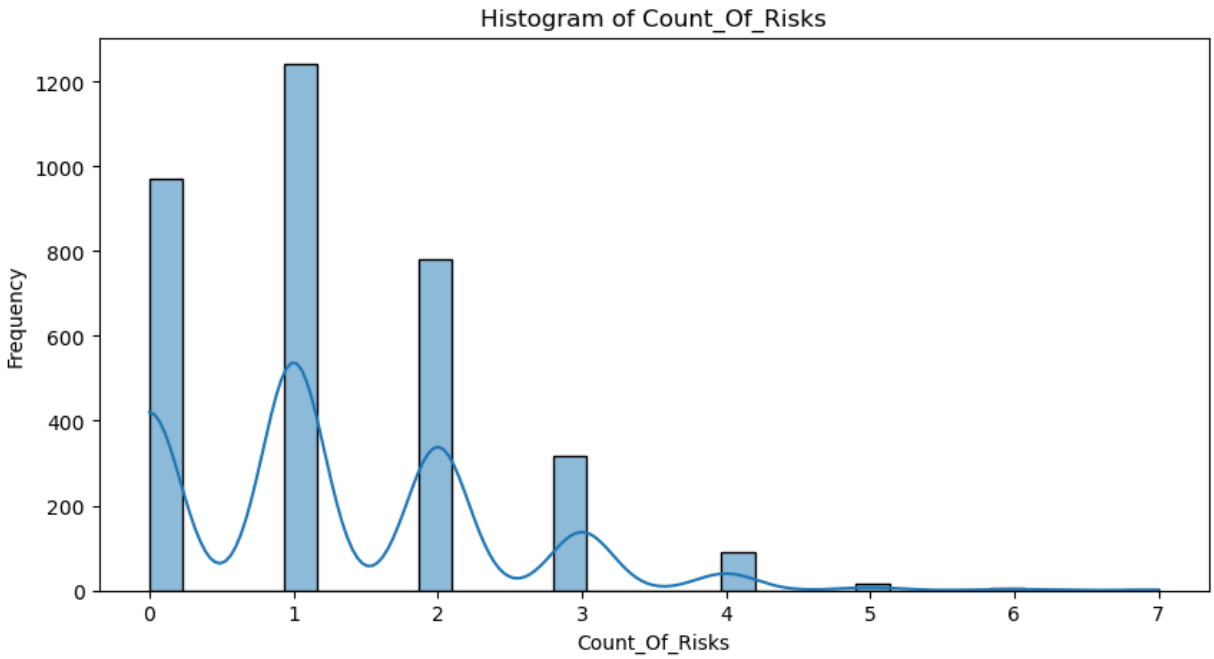
- The log transformation reduces the skewness observed in the original data, but outliers are still present.
- The majority of the data is concentrated below the 1 value, indicating a right-skewed distribution even after log transformation.



This histogram displays the distribution of the **Dexa_Freq_During_Rx** variable.

Key Observations:

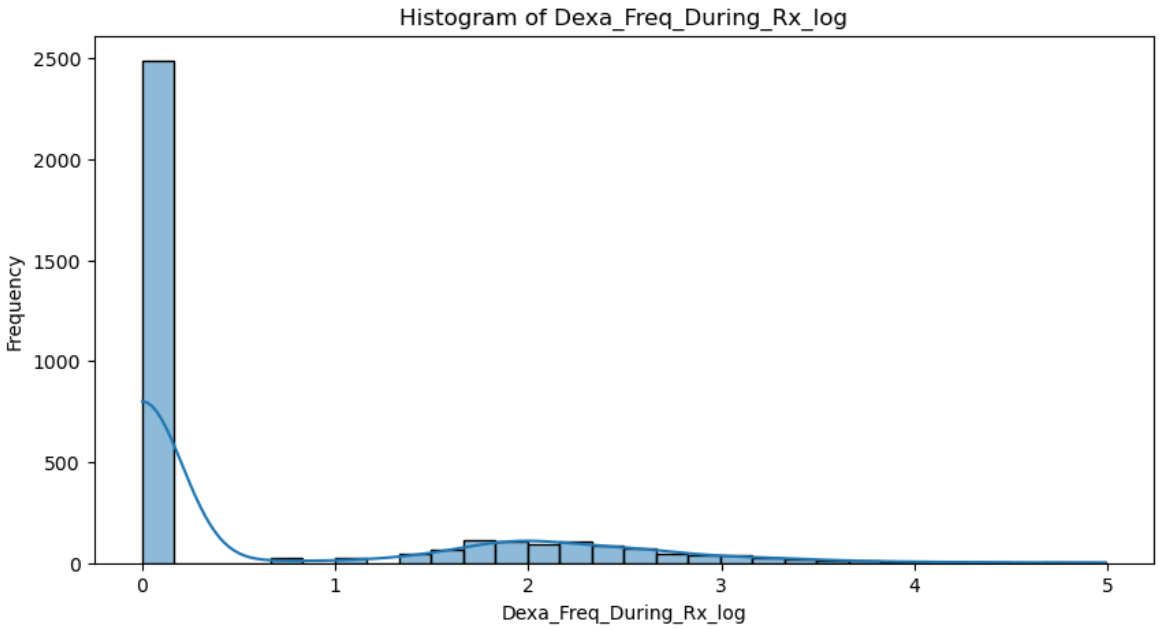
- The distribution is highly right-skewed, with most data points concentrated at the lower end.
- There are a few extreme values extending up to 146, which are potential outliers.



This histogram displays the distribution of the **Count_Of_Risks** variable.

Key Observations:

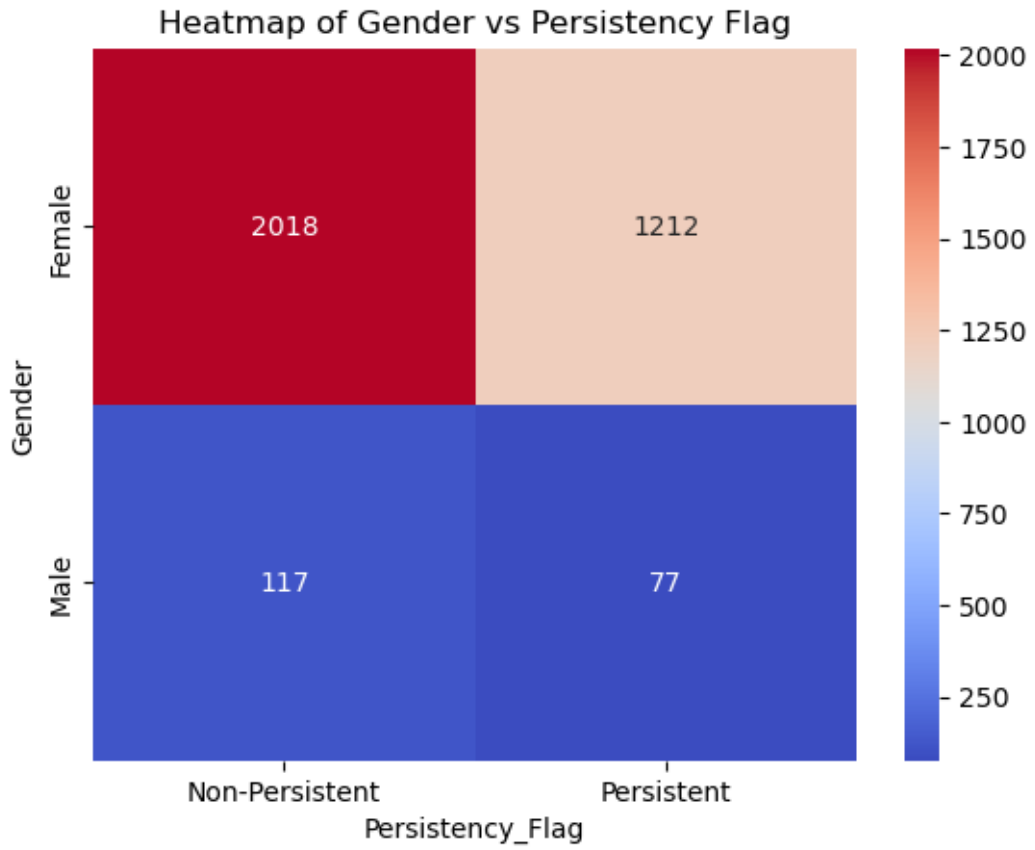
- The distribution shows multiple peaks, indicating several common values within the dataset.
- The majority of the data falls between 0 and 4, with fewer instances of higher risk counts.



This histogram displays the distribution of the log-transformed **Dexa_Freq_During_Rx** variable.

Key Observations:

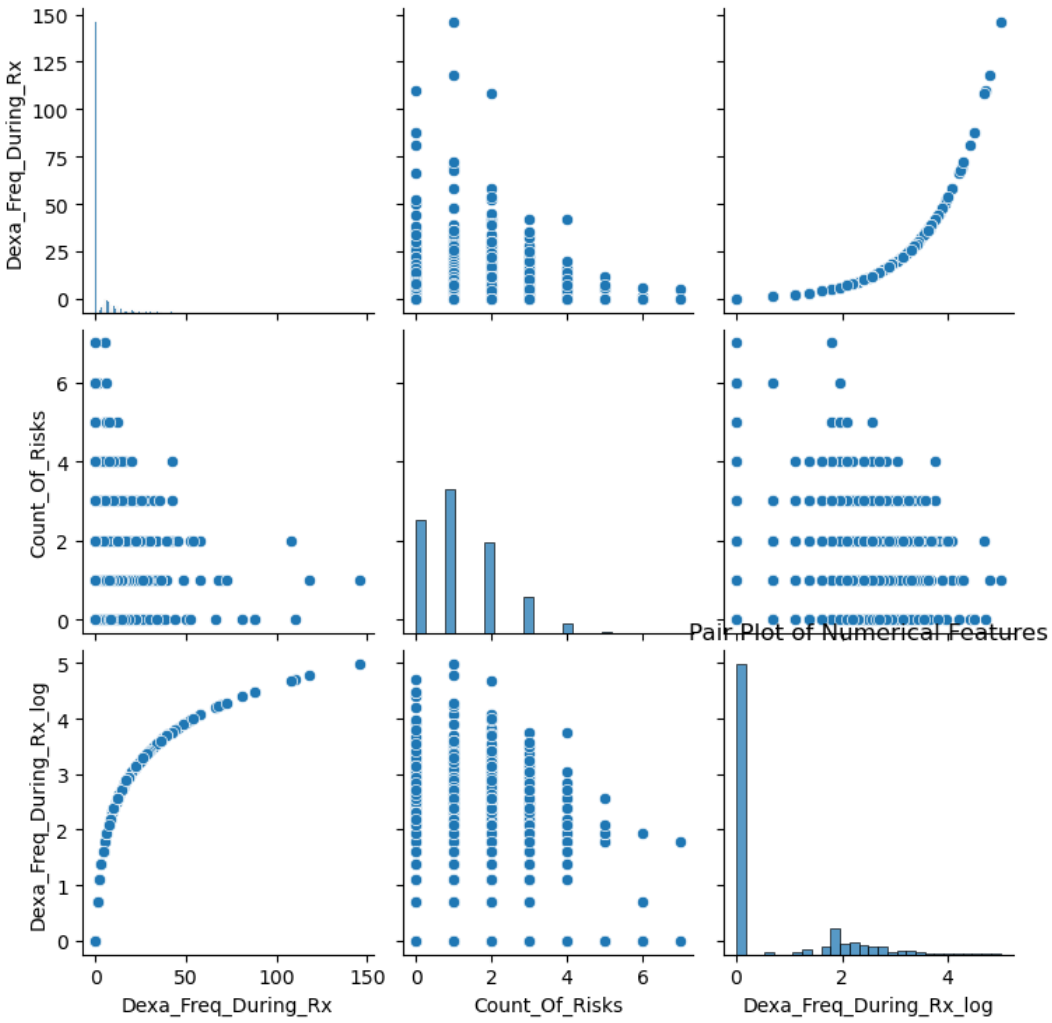
- The log transformation has compressed the extreme values, making the distribution less skewed.
- The majority of the data is still concentrated at the lower end, but the spread is more even compared to the original distribution.



This heatmap visualizes the relationship between **Gender** and **Persistency_Flag**.

Key Observations:

- There is a significantly higher number of females in both persistent and non-persistent categories compared to males.
- The majority of the data points fall into the "Female" category, indicating a gender imbalance in the dataset.
- Among females, non-persistent cases (2018) are much higher compared to persistent cases (1212).
- Among males, non-persistent cases (117) are higher compared to persistent cases (77), but the difference is not as pronounced as in females.



The pair plot visualizes the pairwise relationships between numerical features in the dataset. Each scatter plot represents the relationship between two numerical variables, while the diagonal plots show the distribution of individual variables.

Key Observations:

- **Dexa_Freq_During_Rx vs Count_Of_Risks**: The scatter plot shows a cluster of points at the lower end, indicating that higher counts of risks are associated with higher Dexa frequency.
- **Dexa_Freq_During_Rx vs Dexa_Freq_During_Rx_log**: The scatter plot shows a curved relationship, highlighting the effect of the log transformation in compressing higher values.

- **Count_Of_Risks vs Dexa_Freq_During_Rx_log**: Similar to the original variable, higher counts of risks are associated with higher log-transformed Dexa frequency, but the relationship is more evenly distributed due to the transformation.
- **Individual Distributions**: The diagonal plots show the distributions of individual numerical features.
    - **Dexa_Freq_During_Rx**: Highly right-skewed distribution.
    - **Count_Of_Risks**: Multi-modal distribution with several peaks.
    - **Dexa_Freq_During_Rx_log**: Reduced skewness compared to the original distribution.

Final Recommendations

Based on the findings from the EDA, the following steps are recommended for developing an effective classification model to predict drug persistency:

1. Address Skewness and Outliers:

- Implement log transformation for highly skewed variables like **Dexa_Freq_During_Rx** to normalize the distribution.
- Utilize robust methods or additional transformations to manage the outliers identified in the dataset.

2. Feature Engineering:

- Investigate the creation of new features or combinations of existing ones to capture interactions and non-linear relationships, particularly for variables with weak correlations.
- Apply binning or grouping techniques to numerical variables with a broad range to simplify the data and enhance model performance.

3. Handle Gender Imbalance:

- Address the gender imbalance in the dataset by using stratified sampling or techniques like SMOTE to ensure balanced representation during model training.

4. Dimensionality Reduction:

- Use Principal Component Analysis (PCA) or other dimensionality reduction methods to decrease the number of features while preserving most of the variance. This can help improve model performance and interpretability.

5. Model Selection and Validation:

- Employ cross-validation techniques to assess the performance of various classification models.
- Consider using ensemble methods such as Random Forest or Gradient Boosting to capture complex patterns in the data.