Project : Healthcare - Persistency of a drug

Week 9 Deliverable

Batch: LISUM38

Team: Salus AI Data Avengers				
Name	**Email**	**Country**	**Company**	**Specialisation**
Amrin Shaikh	amrin02@gmail.com	United Kingdom	Freelance	Data Science
Keilor Fallas Prado	kfallasprado@gmail.com	Costa Rica	Immune Technology Institute Madrid	Data Science
Hyejoon Lee	candy0543@gmail.com	United Kingdom	--	Data Science
Raina Singh	rainasinghh1@gmail.com	United States of America	The Ohio State University	Data Science

Problem description:

The primary goal of this project is to analyse patient persistency data and prepare it for predictive modelling. While the dataset has no missing values, it exhibits other quality issues, such as outliers and skewness. These issues are addressed through appropriate data cleansing and transformation techniques to ensure the data is suitable for machine learning.

GitHub Repository Link:

https://github.com/ShaikhAmrin02/Data-Glacier-Internship_project.git

Data Cleansing and Transformation Techniques:

Outlier Handling:

Dexa_Freq_During_Rx (Log Transformation):

- A log transformation was applied to reduce the influence of extreme values.
- The transformation removed 42 outliers, reduced skewness, and aligned the data closer to a normal distribution.

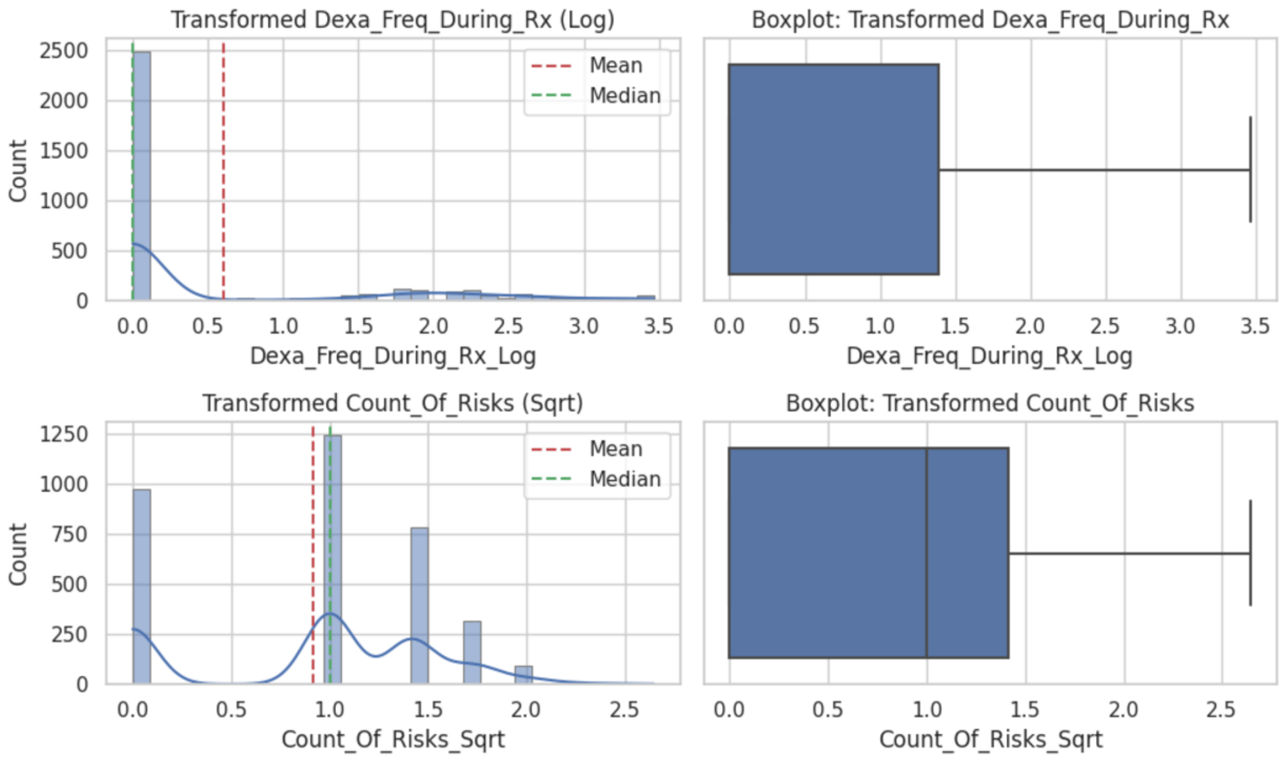
Count_Of_Risks (Z-Score):

- Z-Score method was applied to identify and cap extreme outliers.
- Outliers with Z-scores beyond 3 standard deviations were replaced with the nearest boundary.
- This technique retained most of the data while mitigating the influence of extreme values.

Skewness Handling:

Dexa_Freq_During_Rx: Skewness reduced significantly from 6.81 to 1.33 after log transformation.

Count_Of_Risks: Z-score capping addressed mild outliers, resulting in better data distribution without significant skewness changes.



Visual Insights

- Boxplots and histograms illustrate reduced outliers and improved normality in transformed variables.
- Improved data distribution ensures better model stability and performance.

Summary:

The data cleansing process focused on handling outliers and skewness using two different approaches:

- Log transformation for heavily skewed variables.
- Z-score method for moderate outliers.

These steps ensure that the dataset is ready for robust classification modelling, free from distortions caused by extreme values.