Data Analysis Project Submission

Team Member's Details

Group Name: The Closer

Member 1 Name: Rongala Sreedhar

Member 1 Email: rongalasreedhar@gmail.com

Member 1 Country: Italy

Member 1 College/Company: University of Naples Federico II

Member 1 Specialization: Data Science

Problem Description

The goal is to analyze the 'Healthcare_dataset' to identify data quality issues such as missing values, outliers, and skewness, and to propose appropriate strategies for data cleaning and preprocessing to prepare the dataset for further modeling tasks (e.g., specific classification or regression).

Data Understanding

Dataset: Healthcare_dataset (1).xlsx

Shape: 3424 rows, 69 columns

Type of Data: Structured / Tabular (Healthcare records).

Key Features types:

- Ptid: object
- Persistency_Flag: object
- Gender: object
- Race: object
- Ethnicity: object
- Region: object
- Age_Bucket: object
- Ntm_Speciality: object
- Ntm_Specialist_Flag: object
- Ntm_Speciality_Bucket: object

...(and more)

Problems in the Data

Data Analysis Project Submission

No missing values found.

Outliers detected (using IQR method):

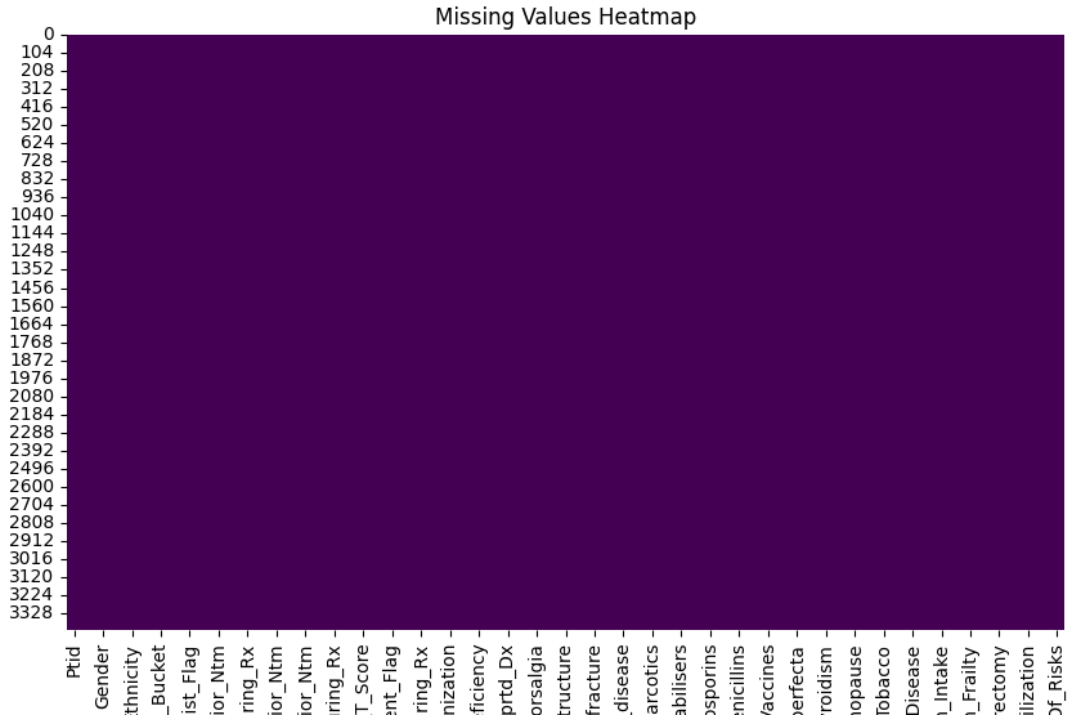
- Dexa_Freq_During_Rx: 460 potential outliers
- Count_Of_Risks: 8 potential outliers

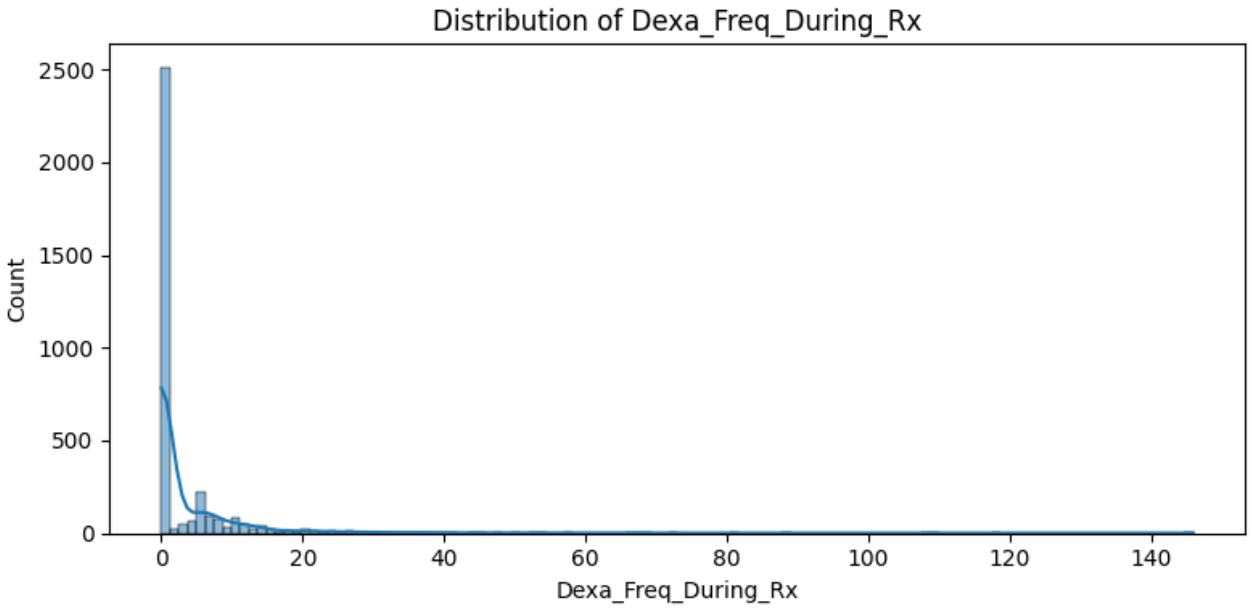
Skewed Features (|skew| > 1):

- Dexa_Freq_During_Rx: 6.81

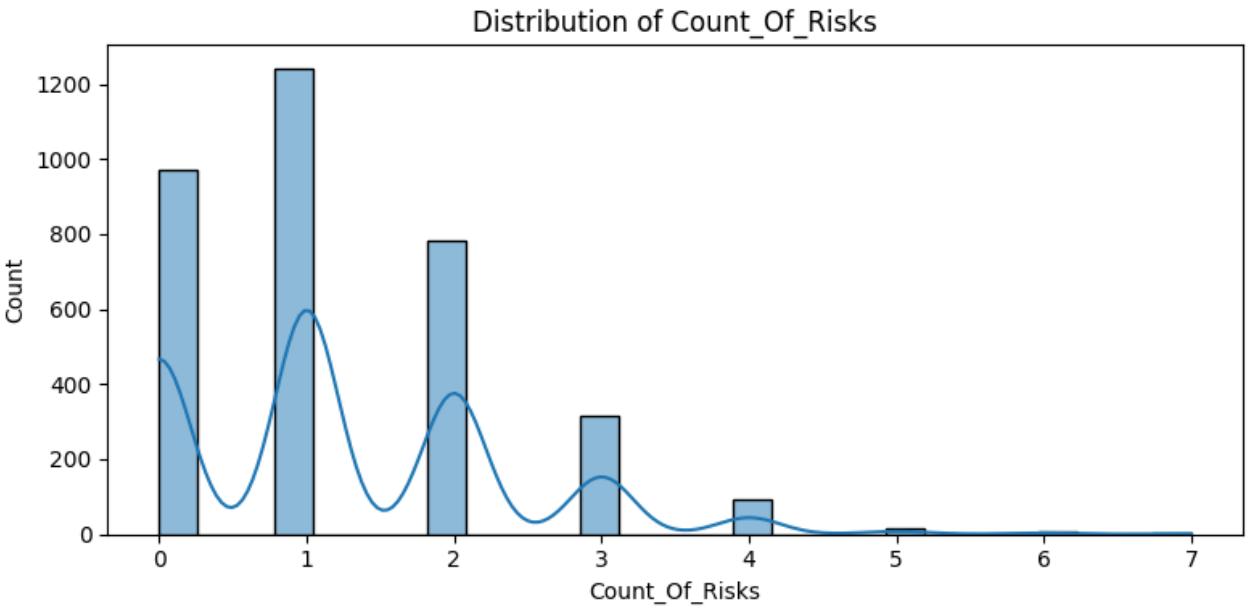
Data Analysis Project Submission

Visualizations





Data Analysis Project Submission



Data Analysis Project Submission

Approaches to Handle Data Problems

1. Missing Values:
   - Numerical: Impute with Median (robust to outliers) or Mean (if normal).
   - Categorical: Impute with Mode or create a 'Missing' category.
   - Rationale: Preserves data volume compared to dropping rows.

2. Outliers:
   - Log Transformation: To reduce the impact of extreme values.
   - Capping (Winsorization): Limiting extreme values to the 1st/99th percentiles.
   - Rationale: Linear models are sensitive to outliers; these methods reduce their influence without losing data points.

3. Skewness:
   - Apply Log or Box-Cox transformations to normalize distributions.
   - Rationale: Many statistical models assume normality.

GitHub Repo Link

https://github.com/example/repo