Team: Shuyi Yu (1 member)
Email: Shuyiyu418@gmail.com
Country: USA
College/Company: Western Governors University
Specialization: Data Science
Deliverable: Week 8

Problem description: The project involves analyzing a dataset to understand its structure, identify potential problems like missing values, outliers, and skewness, and apply appropriate solutions to prepare the data for further analysis or modeling.

This analysis has prepared the dataset by addressing key issues like outliers and skewness. The data is now ready for further analysis or modeling.

Data Understanding:

- **Data Structure**: The dataset contains **3424 entries** and **69 columns**.
- **Data Types**:
    - The majority of the columns are **categorical** (object type).
    - Two columns are **numeric**: Dexa_Freq_During_Rx and Count_Of_Risks (int64 type).
- **No Missing Values**: All columns are non-null.

Categorical Columns and Their Unique Values:

- **Ptid**: 3424 unique values
- **Persistency_Flag**: 2 unique values
- **Gender**: 2 unique values
- **Race**: 4 unique values
- **Ethnicity**: 3 unique values
- **Region**: 5 unique values
- **Age_Bucket**: 4 unique values
- **Ntm_Speciality**: 36 unique values
- **Ntm_Specialist_Flag**: 2 unique values
- **Ntm_Speciality_Bucket**: 3 unique values
- **Gluco_Record_Prior_Ntm**: 2 unique values
- **Gluco_Record_During_Rx**: 2 unique values
- **Dexa_During_Rx**: 2 unique values
- **Frag_Frac_Prior_Ntm**: 2 unique values
- **Frag_Frac_During_Rx**: 2 unique values

- **Risk_Segment_Prior_Ntm**: 2 unique values
- **Tscore_Bucket_Prior_Ntm**: 2 unique values
- **Risk_Segment_During_Rx**: 3 unique values
- **Tscore_Bucket_During_Rx**: 3 unique values
- **Change_T_Score**: 4 unique values
- **Change_Risk_Segment**: 4 unique values
- **Adherent_Flag**: 2 unique values
- **Idn_Indicator**: 2 unique values
- **Injectable_Experience_During_Rx**: 2 unique values
- **Comorb_Encounter_For_Screening_For_Malignant_Neoplasms**: 2 unique values
- **Comorb_Encounter_For_Immunization**: 2 unique values
- **Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx**: 2 unique values
- **Comorb_Vitamin_D_Deficiency**: 2 unique values
- **Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified**: 2 unique values
- **Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx**: 2 unique values
- **Comorb_Long_Term_Current_Drug_Therapy**: 2 unique values
- **Comorb_Dorsalgia**: 2 unique values
- **Comorb_Personal_History_Of_Other_Diseases_And_Conditions**: 2 unique values
- **Comorb_Other_Disorders_Of_Bone_Density_And_Structure**: 2 unique values
- **Comorb_Disorders_of_Lipoprotein_Metabolism_and_Other_Lipidemias**: 2 unique values
- **Comorb_Osteoporosis_Without_Current_Pathological_Fracture**: 2 unique values
- **Comorb_Personal_History_of_Malignant_Neoplasm**: 2 unique values
- **Comorb_Gastroesophageal_Reflux_Disease**: 2 unique values
- **Concom_Cholesterol_And_Triglyceride_Regulating_Preparations**: 2 unique values
- **Concom_Narcotics**: 2 unique values
- **Concom_Systemic_Corticosteroids_Plain**: 2 unique values
- **Concom_Anti_Depressants_And_Mood_Stabilisers**: 2 unique values
- **Concom_Fluoroquinolones**: 2 unique values
- **Concom_Cephalosporins**: 2 unique values
- **Concom_Macrolides_And_Similar_Types**: 2 unique values
- **Concom_Broad_Spectrum_Penicillins**: 2 unique values
- **Concom_Anaesthetics_General**: 2 unique values
- **Concom_Viral_Vaccines**: 2 unique values
- **Risk_Type_1_Insulin_Dependent_Diabetes**: 2 unique values
- **Risk_Osteogenesis_Imperfecta**: 2 unique values

- **Risk_Rheumatoid_Arthritis**: 2 unique values
- **Risk_Untreated_Chronic_Hyperthyroidism**: 2 unique values
- **Risk_Untreated_Chronic_Hypogonadism**: 2 unique values
- **Risk_Untreated_Early_Menopause**: 2 unique values
- **Risk_Patient_Parent_Fractured_Their_Hip**: 2 unique values
- **Risk_Smoking_Tobacco**: 2 unique values
- **Risk_Chronic_Malnutrition_Or_Malabsorption**: 2 unique values
- **Risk_Chronic_Liver_Disease**: 2 unique values
- **Risk_Family_History_Of_Osteoporosis**: 2 unique values
- **Risk_Low_Calcium_Intake**: 2 unique values
- **Risk_Vitamin_D_Insufficiency**: 2 unique values
- **Risk_Poor_Health_Frailty**: 2 unique values
- **Risk_Excessive_Thinness**: 2 unique values
- **Risk_Hysterectomy_Oophorectomy**: 2 unique values
- **Risk_Estrogen_Deficiency**: 2 unique values
- **Risk_Immobilization**: 2 unique values
- **Risk_Recurring_Falls**: 2 unique values

Data Issues and Problem Identification:

The dataset contains both numeric and categorical columns. **Numeric columns** require further analysis to identify issues such as:

- **Missing Values**: There are **no missing values** in the dataset.
- **Outliers**: Outliers were identified in columns such as Dexa_Freq_During_Rx.
- **Skewness**: Skewness was found in numeric columns, particularly in Dexa_Freq_During_Rx (skewness = 1.33).

Solutions Applied:

- **Outliers**: Outliers were addressed using the **IQR method** to cap or remove them, ensuring they don't unduly influence the analysis.
- **Skewness**: A **log transformation** was applied to highly skewed columns such as Dexa_Freq_During_Rx to normalize the data distribution.
- **No Missing Values**: Although there were no missing values, various techniques like filling NA values with medians were considered if needed in the future.

Skewness and Outliers:

- **Log Transformation Applied to Highly Skewed Columns**: Dexa_Freq_During_Rx.
- **Skewness of Numeric Columns**:
  - Dexa_Freq_During_Rx: 1.33 (log transformation applied)
  - Count_Of_Risks: 0.78

Significant Results:

- **Chi-square Test for Region vs. Persistency_Flag**:
  - **p-value**: *4.48×10−64.48 \times 10^{-6}*4.48×10−6
  - **Conclusion**: There is a significant relationship between **Region** and **Persistency_Flag** since the p-value is much lower than the 0.05 significance level.