Team: Shuyi Yu (1 member)
Email: Shuyiyu418@gmail.com
Country: USA
College/Company: Western Governors University
Specialization: Data Science
Deliverable: Week 9
Problem Description: This project aims to analyze healthcare persistency data to determine factors that influence whether patients remain persistent or non-persistent in their treatment regimens. The dataset includes various columns, such as patient demographic details, risk factors, comorbidities, treatment characteristics, and categorical and numeric variables. The primary goal of the project is to thoroughly clean and transform the dataset to prepare it for further analysis and modeling, addressing issues like missing values, outliers, and categorical encoding.

Data Cleansing and Transformation:

1. **Handling Missing Values**: The dataset was examined using the isnull() function. While no explicit NaN values were present, certain fields contained placeholder values like 'Unknown' or 'N/A', which were treated as missing values. These were imputed where applicable using mode or frequent value techniques.
2. **Outlier Detection and Handling**: Outliers were identified in numeric columns using the Interquartile Range (IQR) method. Values outside 1.5 times the IQR were flagged as potential outliers and reviewed for further treatment, ensuring accurate downstream analysis.
3. **Categorical Encoding**:
    a. **Binary Columns**: Columns with 'Y'/'N' values were converted into binary (1/0) encoding for compatibility with modeling.
    b. **One-Hot Encoding**: Text columns, such as 'Race', 'Region', and 'Ntm_Speciality', were one-hot encoded to allow for proper representation in models.
    c. **Ordinal Columns**: Columns with ordered categories, such as 'Age_Bucket', 'Tscore_Bucket_Prior_Ntm', and 'Tscore_Bucket_During_Rx', were mapped to ordinal numeric values to maintain the order for analysis.
4. **NLP Cleaning**: Text-based columns were cleaned using regex and Python techniques. This involved removing unwanted characters, punctuation, and digits, and standardizing text to lowercase. These steps ensured that the text data was ready for NLP-based modeling and analysis.
5. **Data Validation and Transformation**: After all cleaning steps, the transformed dataset was validated to ensure that all columns were either numeric or properly encoded. This validation step guaranteed that the data was in the correct format for subsequent modeling.