Week 8 Deliverable: Healthcare Project
Group Name: The Data Doctors

Data Understanding:

- The dataset provides the factors impacting the patient's persistence to an ABC pharmaceutical company's drugs prescribed by various NTM specialists.
- The aim is to build a machine-learning model that classifies the patients into Persistent and Non-persistent.
- The dataset contains 69 features that are divided into various categories of features:
    - 1 Target feature - Persistency_Flag
    - 1 unique identifier for each patient - Ptid
    - 6 Demographics of the patient - Age_Bucket, Gender, Race, Ethnicity, Region, Idn_Indicator
    - 3 physician specialty attributes - Ntm_Speciality, Ntm_Specialist_Flag, Ntm_Specialist_Bucket
    - 13 Clinical factors - Tscore details, Risk_Segment details, Multiple risk factors count, DEXA details, Fragility fracture details, Glucocorticoid details
    - 45 Disease/Treatment factors - Injectable drugs, Risk factors, Comorbidities, Concomitancies, Adherence to therapy
- The total number of records is 3425.
- There are no missing values in the dataset (other than 'unknown'). Hence, there is no need to handle them.

Type of Data:

- The dataset consists of a high majority of categorical data rather than numerical data.
- Among the given features, 68 are independent variables and the target feature is the Persistency_Flag.

Outliers Detection:

- Z-score
- Boxplot
- Inter Quantile Range(IQR)
- Histogram (detecting skewness)

Handling Outliers:

- Winsorize method

- Log transformation
- Median Absolute Deviation (MAD) method
- Box-Cox Transformation
- Square-root Transformation
- Inverse Transformation

Missing Values:

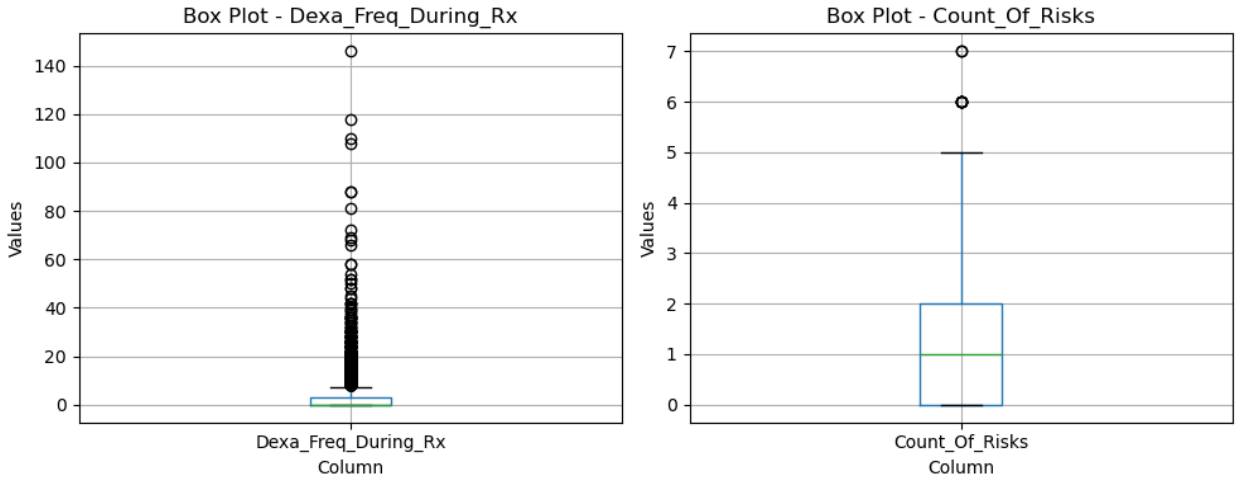
- No missing values found at this stage

```
df.isna().any().sum()
[81]  ✓  0.0s
...   0
```

Numerical Variables:

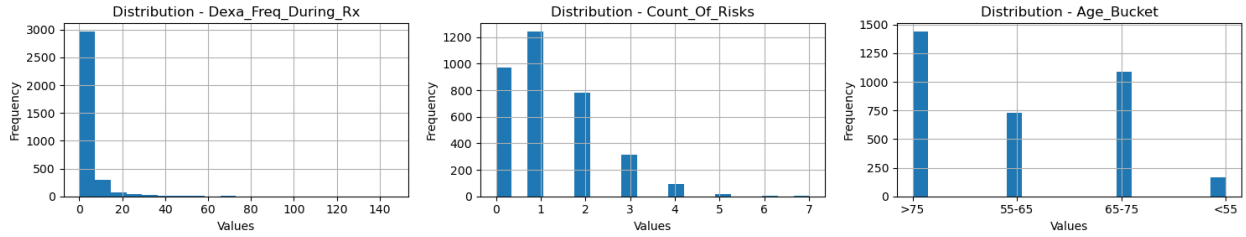
- Dexa_Freq_During_Rx
- Count_Of_Risks
- Age_Bucket (doubtful)

Outliers:

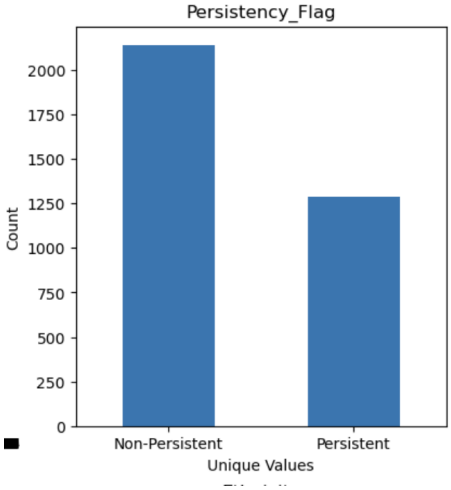


- We can see both columns have outliers
- Data is positively skewed

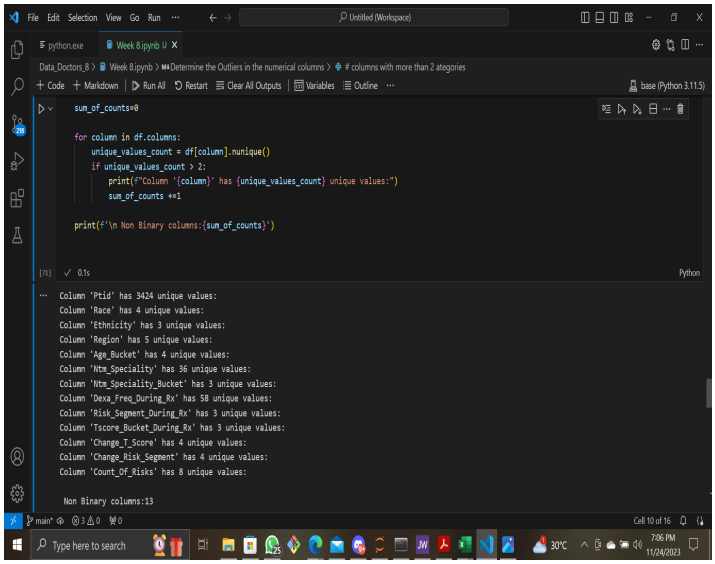
Distribution:



Target variable value count for both categories:

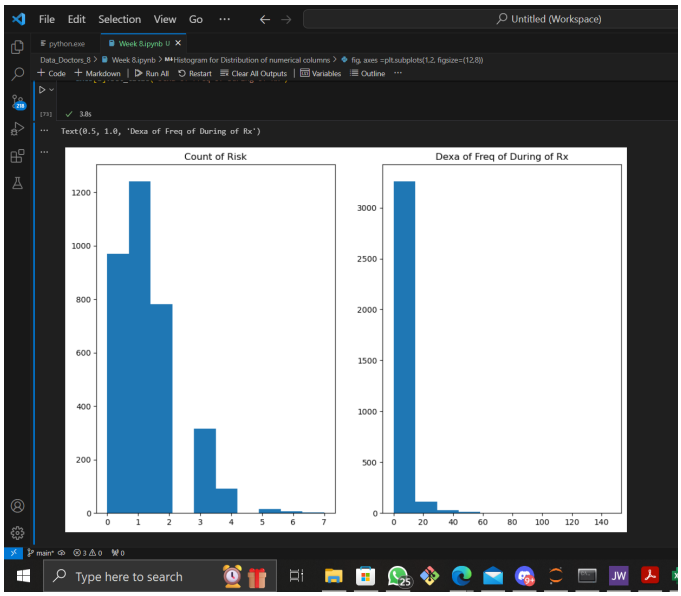


All the 68 columns have nominal categories and 13 out of them are non-binary categories. We have the options, like one-hot, frequency encoding and more, for encoding the data but as some have more than one categories we cannot use one-hot only. We'll either use in combination another.



- Patient ID will mostly be dropped in the long run
- Ntm_Speciality has 36 unique values, we want reduce it to just two: Generalist and Specialist
- Count of Risks has 8 unique values, we might want to reduce to 0, 1, 2, 3, >3
- Dexa_Freq_During_Rx has 58 unique values, we might want to reduce it to 0-6, 6-12, 12-18, 18- 4, 24-30 and >30

Reducing the bins to 5 makes more visual sense in my opinion



- There some unknown in the Risk Segment During prescription column, we might want to input the value from prior column to during column as about 86% considering that about 86% recorded not change in value

No change	Count of No change	% Change
Improved	94	4.880582
No change	1659	86.13707
Worsened	173	8.982347
Total	1926	