Data Glacier

Your Deep Learning Partner

Week 9 Deliverables

Group Name: The Data Doctors

25-Nov-2023

Team Details

Name: Noah Gallego Email: noahgallego394@gmail.com Country: United States College/Company: California State University Bakersfield Specialization: Data Science	Name: Tomisin Abimbola Adeniyi Email: tomisin_adeniyi11@yahoo.com Country: Nigeria College/Company: N/A Specialization: Data Science
Name: Mohammad Shehzar Khan Email: mshehzarkhan@gmail.com Country: Turkey College/Company: Koç University Specialization: Data Science	Name: Ashish Sasanapuri Email: sashrao21@gmail.com Country: India College/Company: N/A Specialization: Data Science

Problem Description

One challenge for all Pharmaceutical companies is to understand the persistence of a drug as per the physician's prescription. To solve this problem ABC Pharma company approached an analytics company to automate this process of identification.

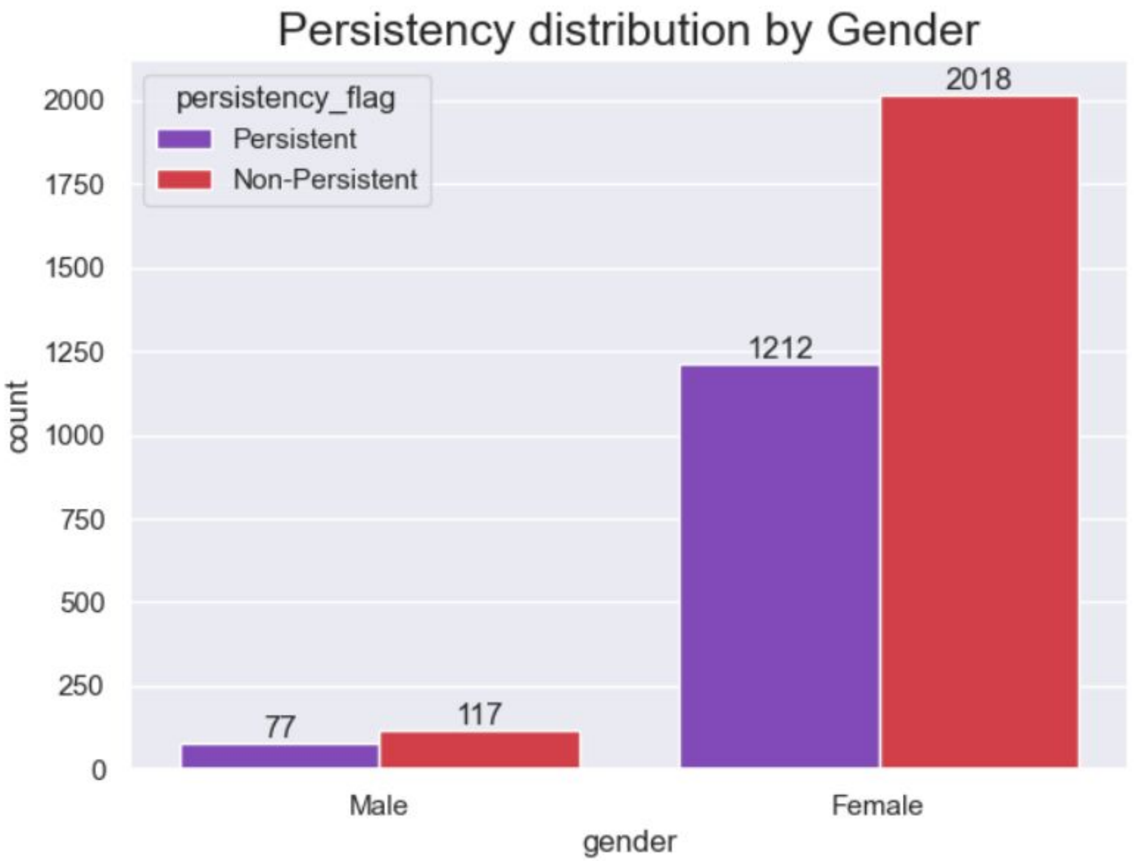
Individual Work by Ashish

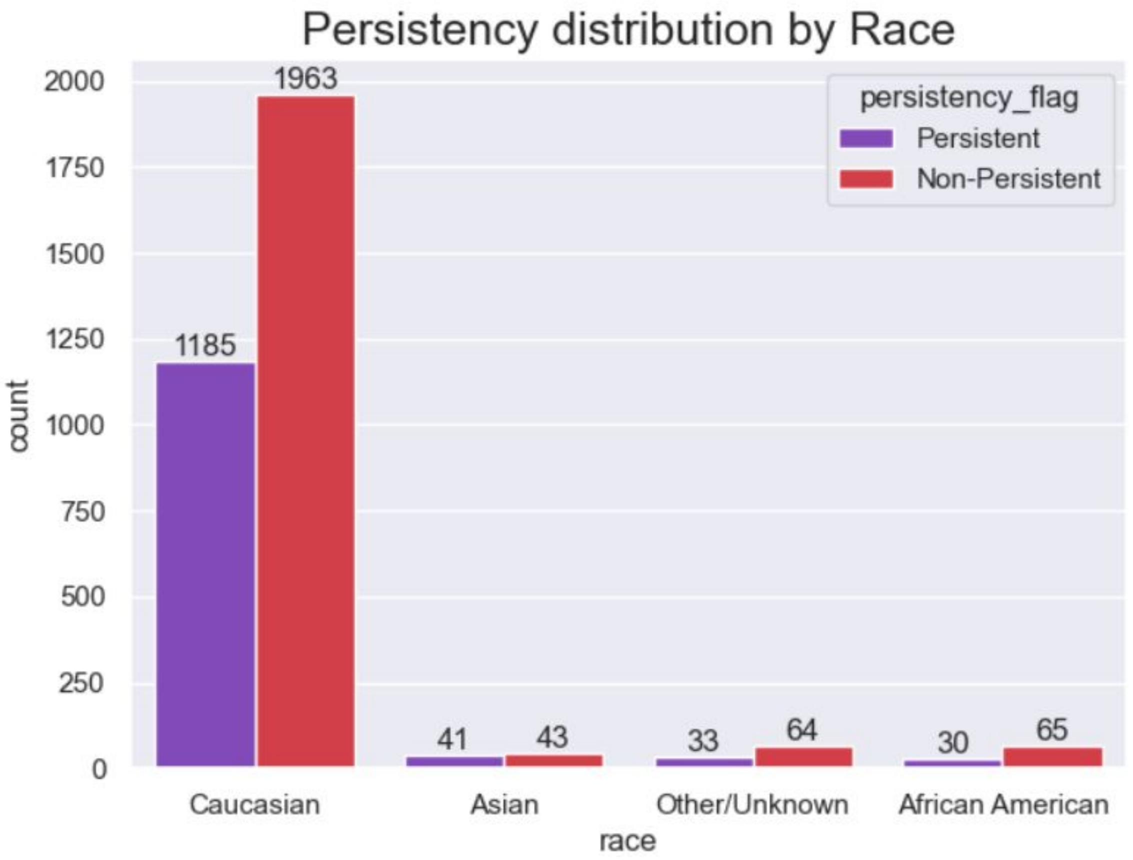
Data Glacier

Your Deep Learning Partner

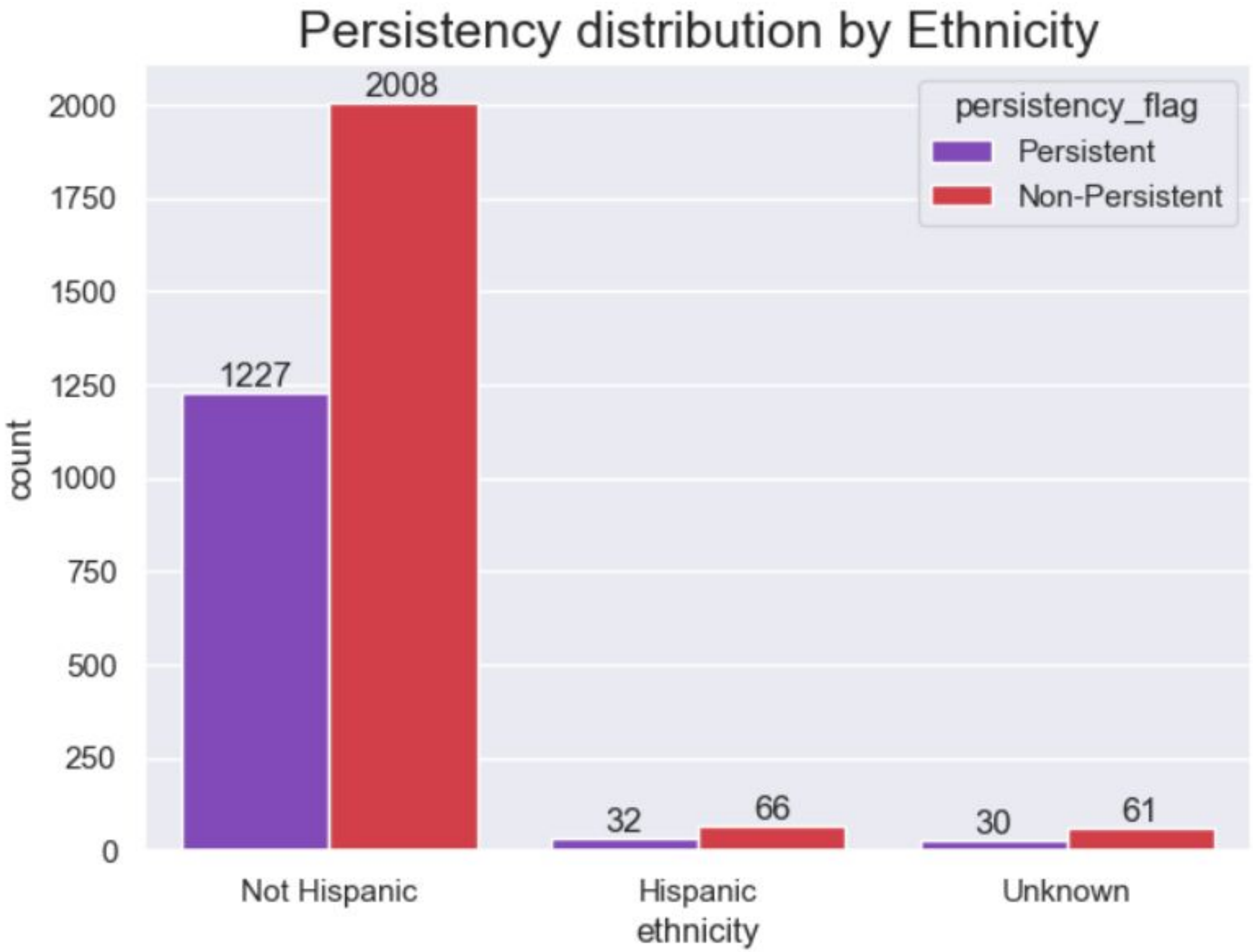
Data Preprocessing

- Demographics such as *Gender*, *Race*, and *Ethnicity* features can be dropped from the dataset as they might introduce bias in the model due to bias in values in their respective features.



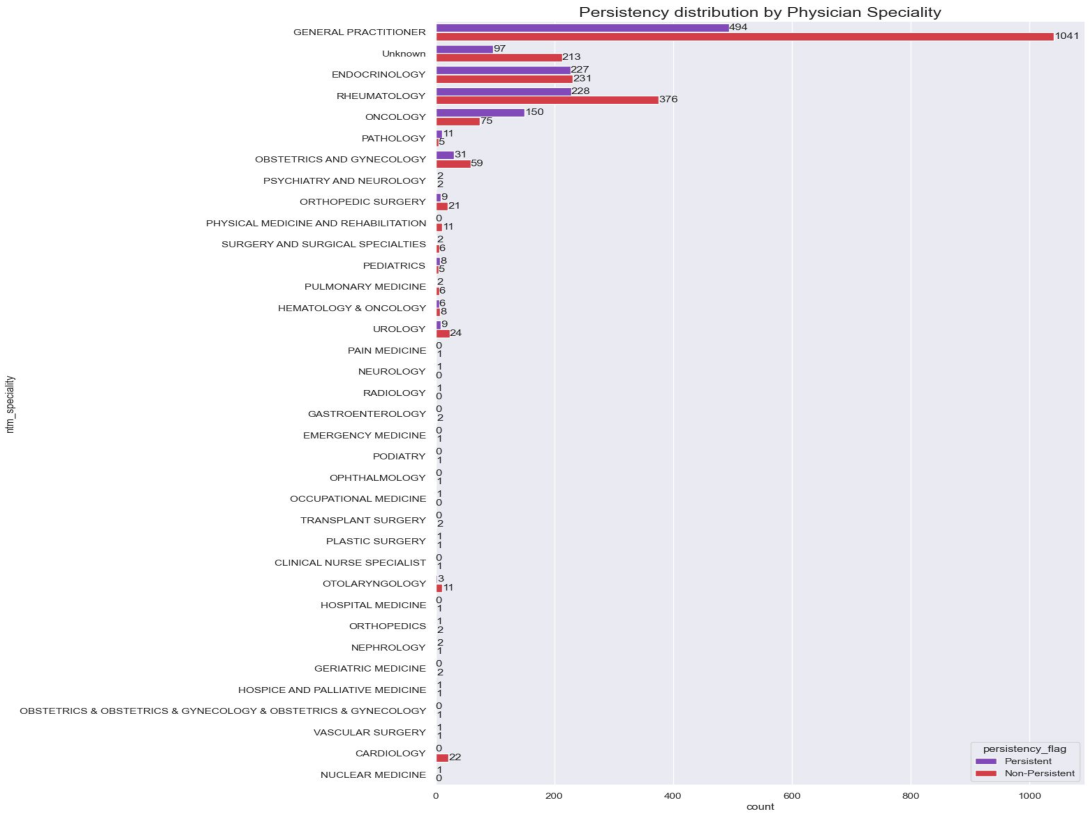


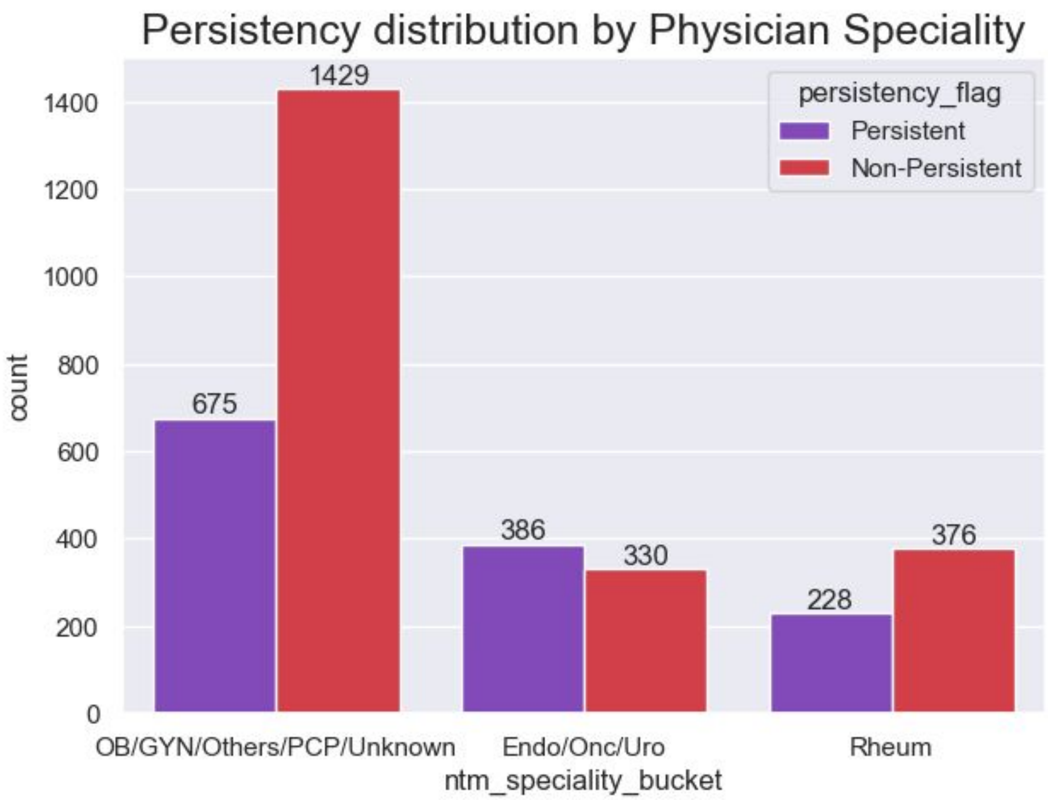
Data Preprocessing(*cont.*)



Data Preprocessing(*cont.*)

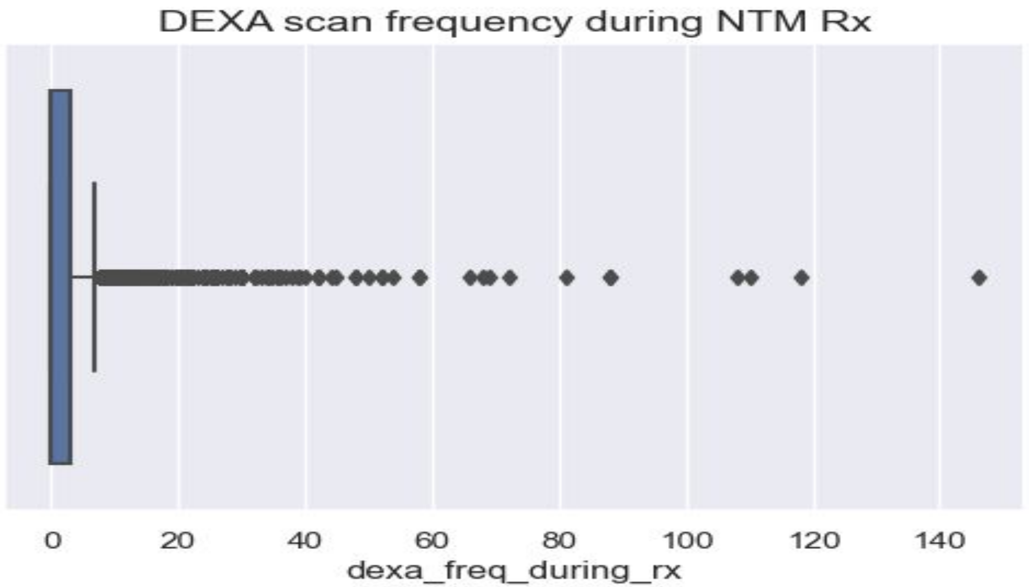
- NTM Physician attributes such *ntm_speciality*, and *ntm_speciality_bucket* features provide the same information but the former contains outliers. Hence, we can drop the former feature and keep *ntm_speciality_bucket feature.*

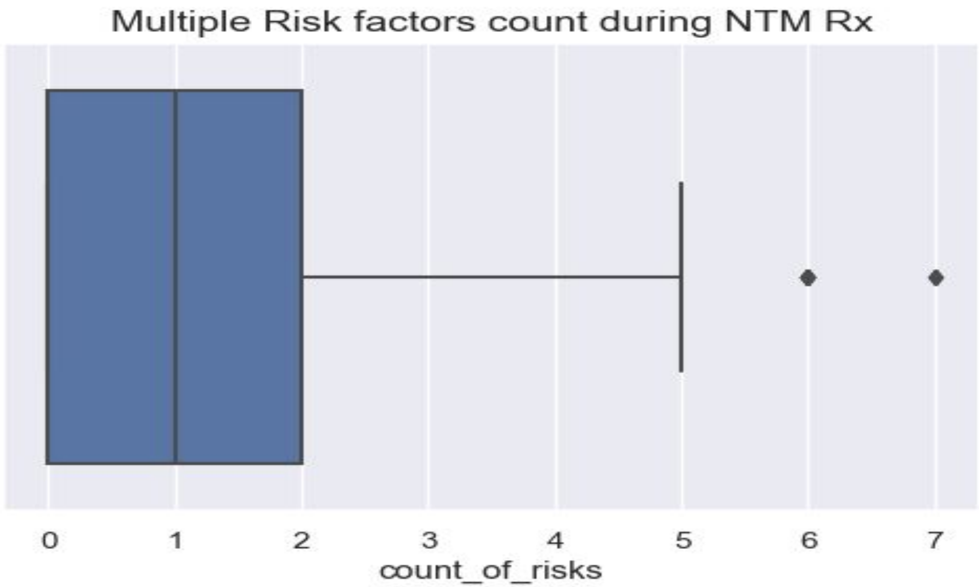




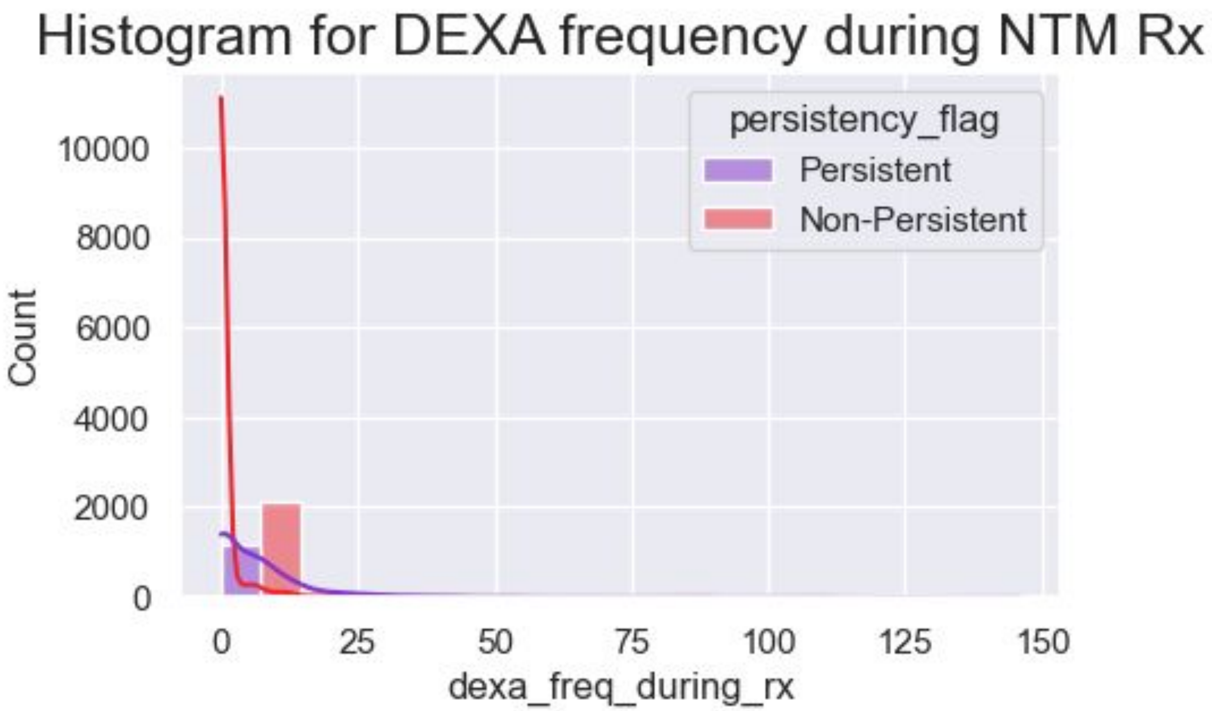
Data Preprocessing(*cont.*)

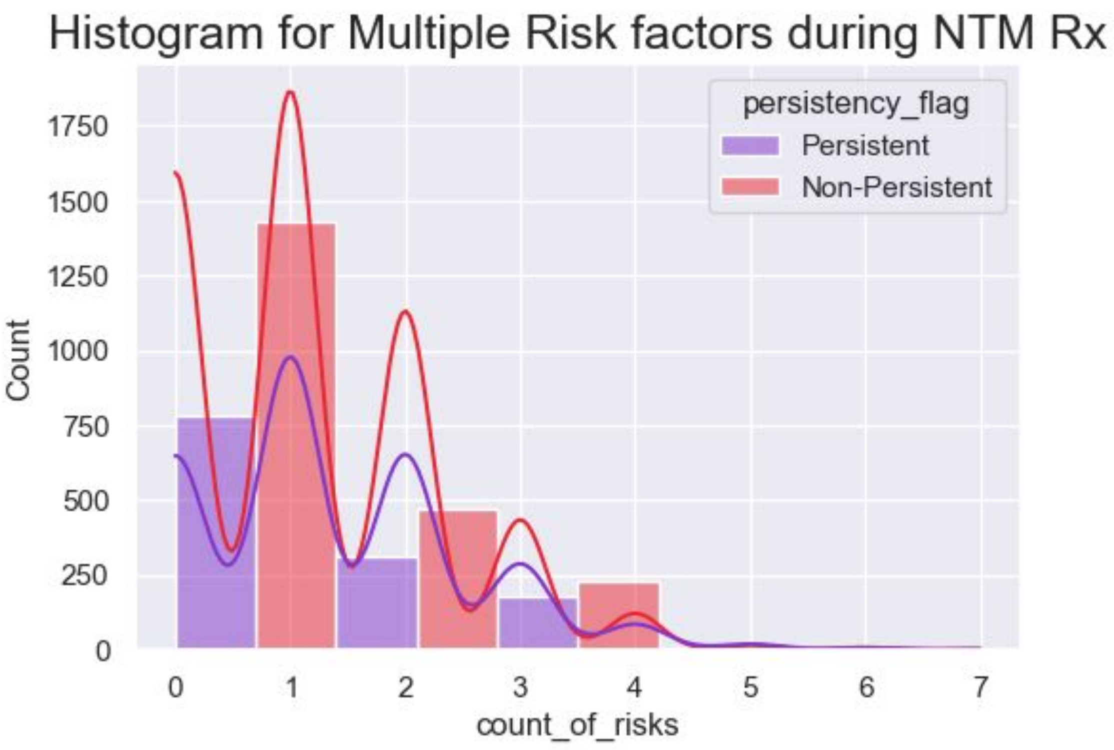
- Outliers are usually observed in Numerical features. 2 of these features in the dataset are - *dexa_freq_during_rx* and *count_of_risks*.
- For the *dexa_freq_during_rx* feature, 4 outlier detection methods were performed -
  - Boxplot visualisation
  - Histogram
  - InterQuartile Range (IQR)
  - Z-Score





Data Preprocessing(*cont.*)





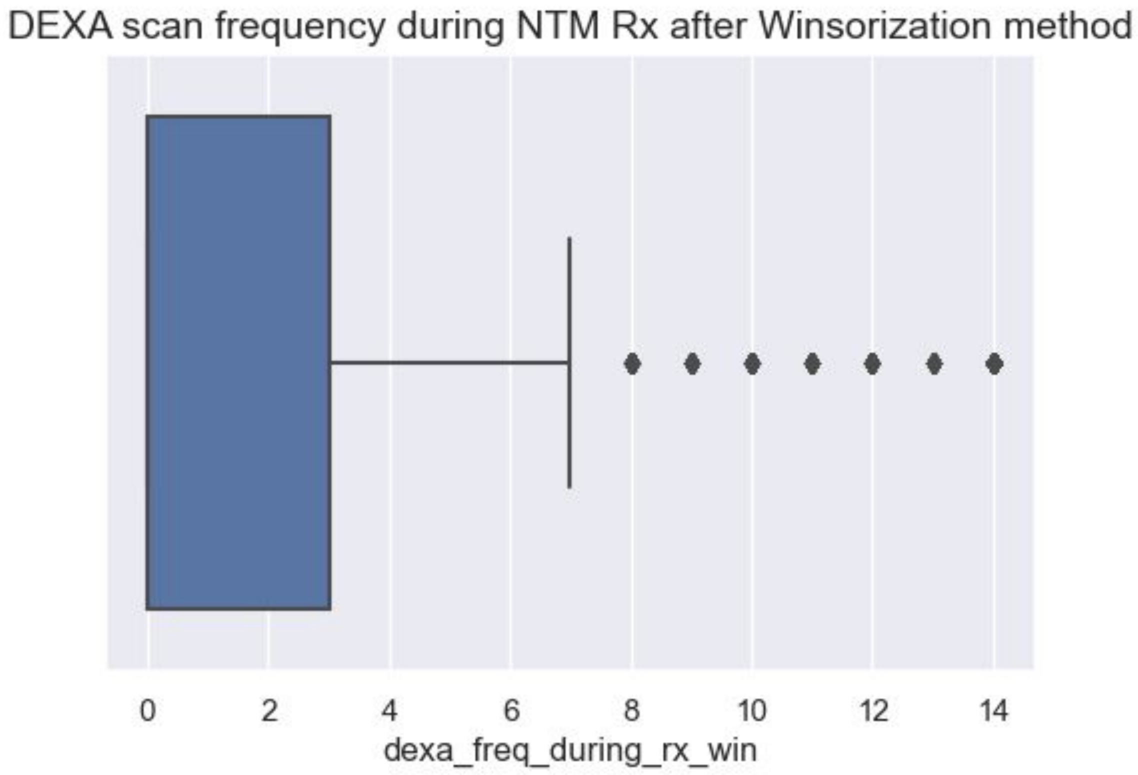
Data Preprocessing(*cont.*)

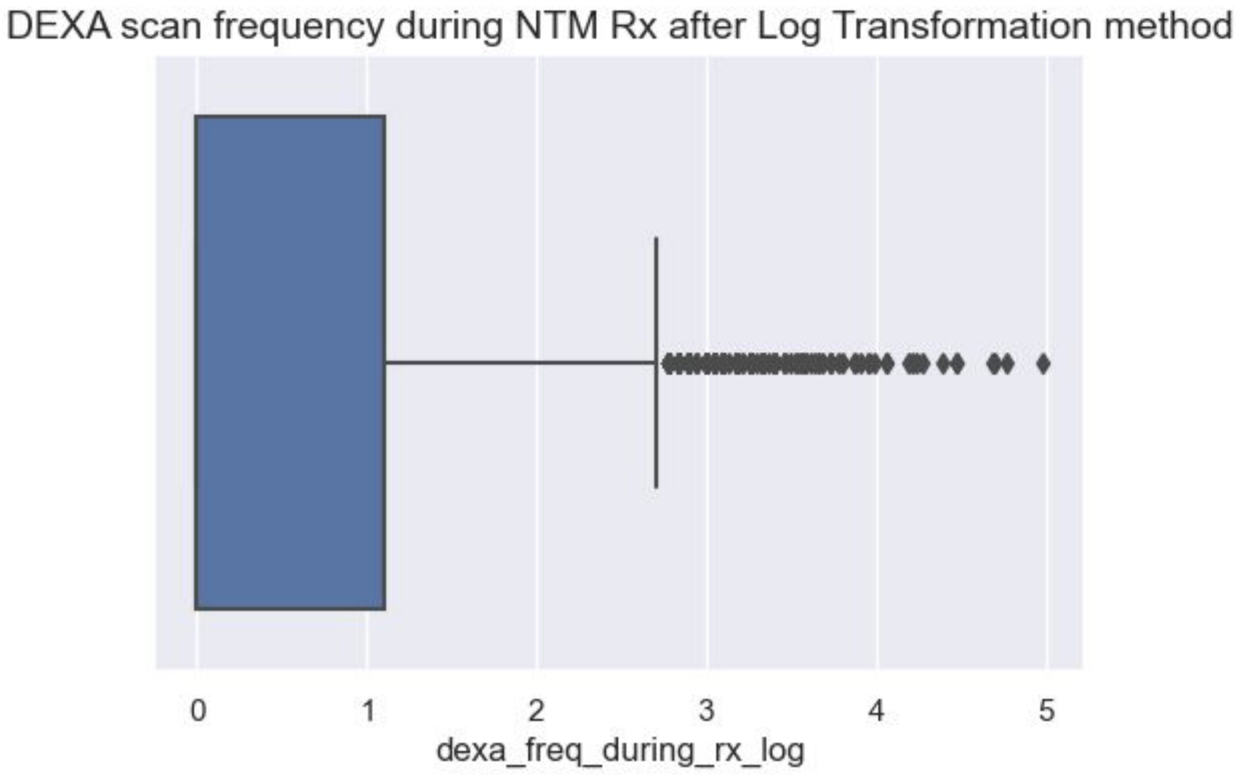
<u>Handling the Outliers</u>:

- For *dexa_freq_during_rx* feature -
  - On calculating **IQR** for this feature, we observed a lower and upper limit of -4.5 and 7.5 respectively.
  - We also applied **Z-Score** for detecting outliers based on a threshold of *1.96*, a value upto which corresponds to *95%* of the data. Total of **119** outliers were observed.
  - Applied **Winsorization** considering the **IQR** detection with lower and upper limits of 5th and 95th percentiles. It reduced outliers and skewness in data from *6.8* to *1.7*.
  - Applied **Log Transformation** but didn't provide significant results as compared to **Winsorization**.
  - Applied **Winsorization** on outliers detected from **Z-Score**. It gave better results than **Log Transformation** but results didn't improve compared to **Winsorization** on IQR.

Data Preprocessing(*cont.*)

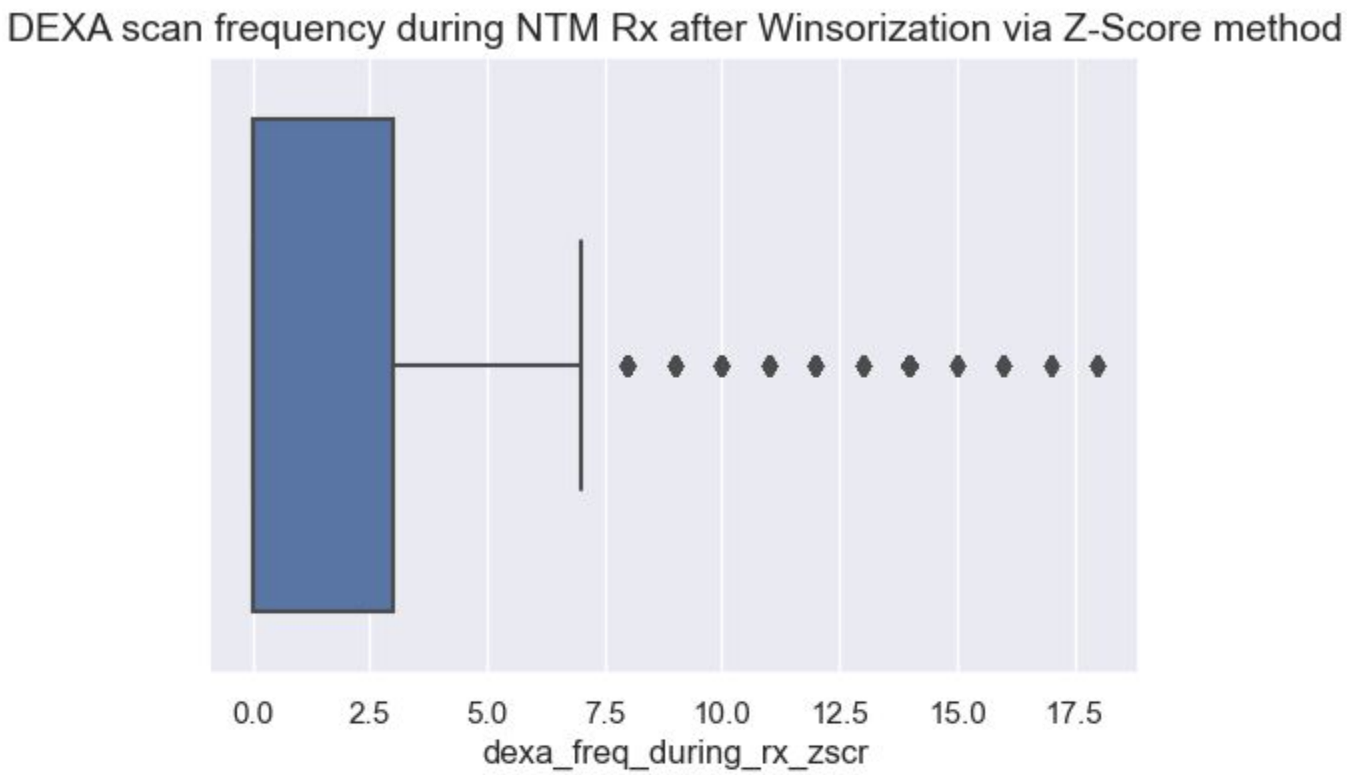
Visualization post outlier treatment:





Data Preprocessing(*cont.*)

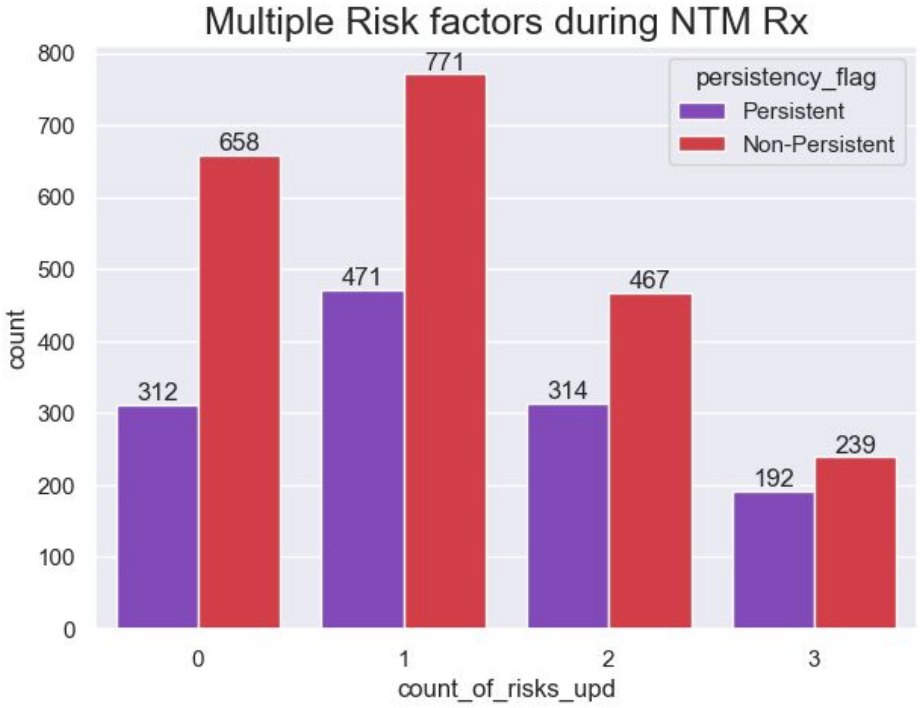
<u>Visualization post outlier treatment:</u>



Data Preprocessing(*cont.*)

<u>Handling the Outliers</u>:

- For *count_of_risks* feature -
    - Only **Boxplot** and **Histogram** plots were plotted for detecting outliers.
    - Based on the distribution of the data in this feature that contains 7 different categories, the approach of reducing the categories to 0, 1, 2, and >3 was employed.



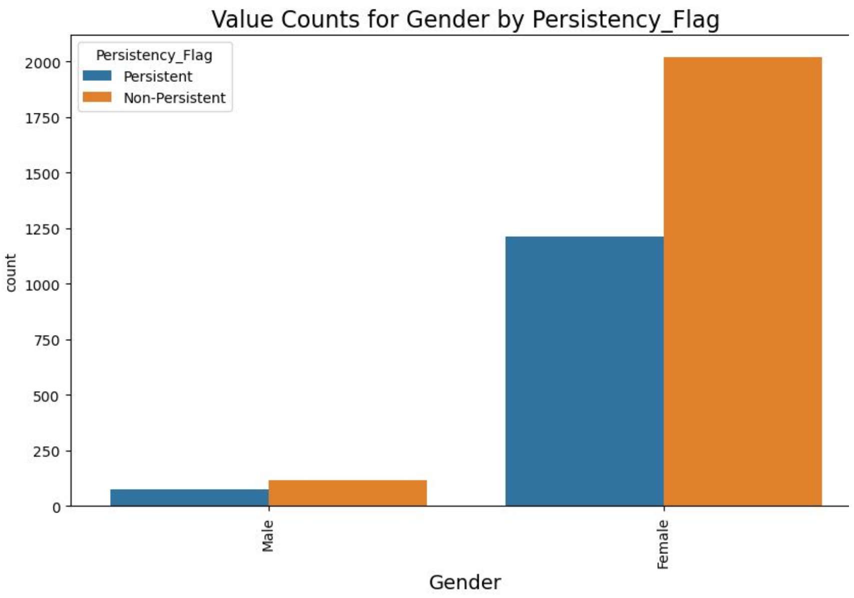
Individual Work by Mohammad

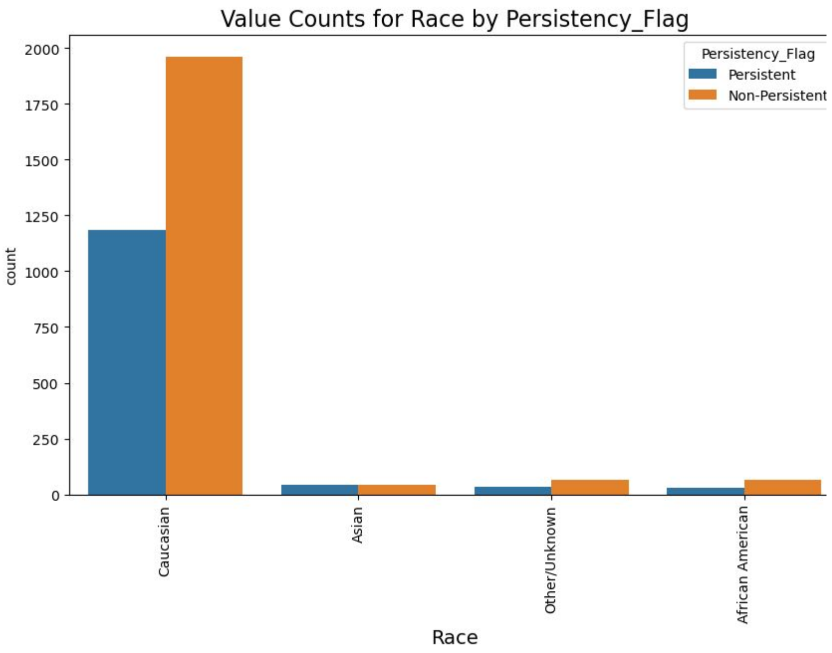
Data Glacier

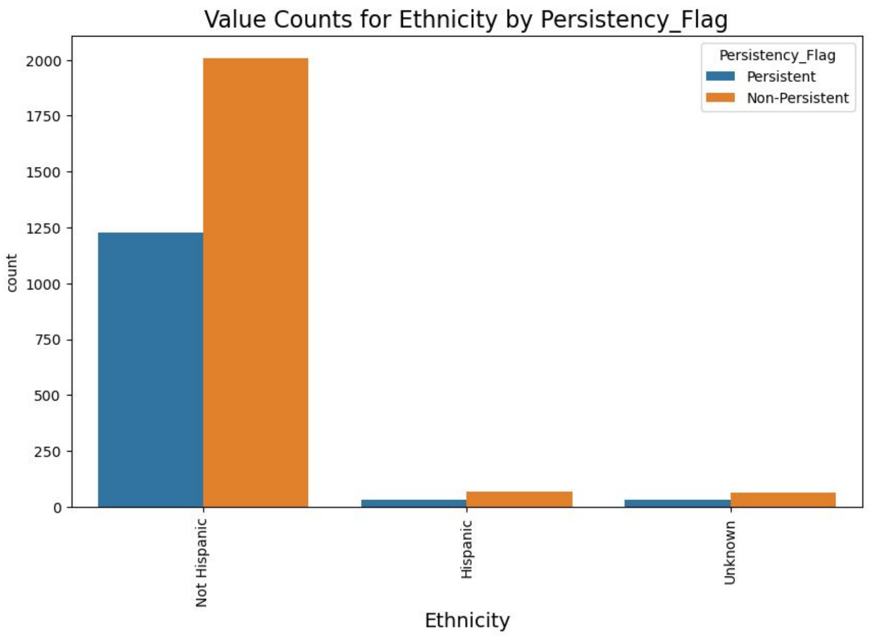
Your Deep Learning Partner

Data Preprocessing

- **Dropping Unnecessary features:** On looking at the value counts of each column by Persistency_Flag, we found some features to be too imbalanced, i.e., the data in one of the categories was negligible compared to the other, and therefore, it doesn't make sense to include those features as they don't add much information, and won't be helpful in ML model training. If we drop such features, we can reduce the complexity of the model.



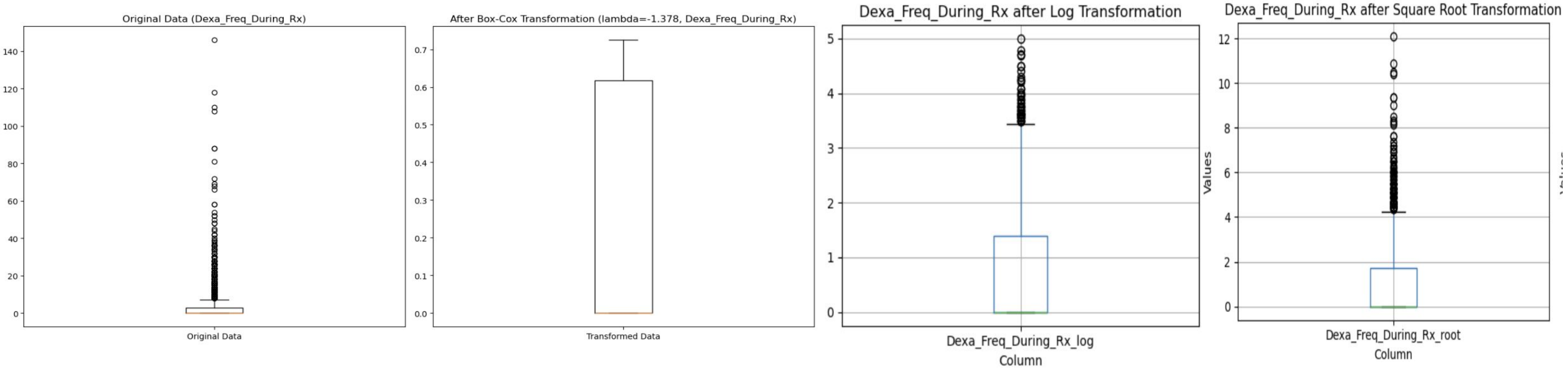




- Some of the features with imbalanced data are shown above.

Data Preprocessing(cont.)

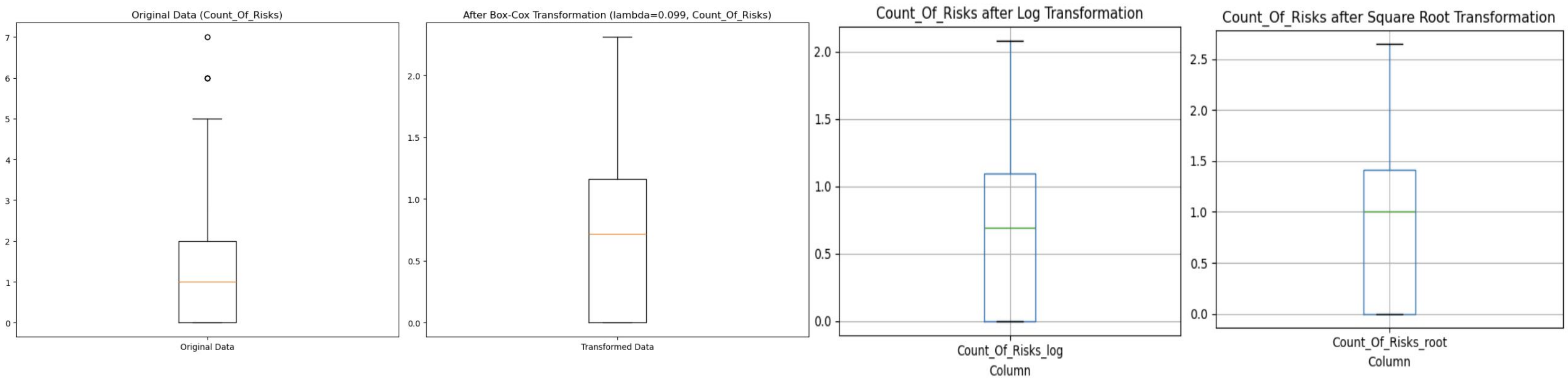
- **Outliers Detection and Handling**: 2 features contain outliers. To deal with them different approaches were tried such as Log-Transformation, Square-Root Transformation, and Box-Cox Transformation.
- Below is the figure for **Box-Cox, Log,** and **Square root Transformations** on **Dexa_Freq_During_Rx** variable.



Here, we can see that box-cox transformation performed the best in handlings outliers, we can proceed with it here.

Data Preprocessing(cont.)

- Below is the figure for **Box-Cox, Log, and Square Root Transformations** on **Count_Of_Risks** variable.



Here, we can say all the techniques performed well. There were not many significant outliers here, so maybe we can leave it as it is, or we can choose one of them, as at least in this way we can scale down the data, which can actually be helpful while training the ML model, as variables with bigger range of values can influence the model more.

Data Preprocessing(cont.)

- **Missing/NA Values**: We don't have any missing values in this data.

```
In [158]: # checking for null values
          df.isna().any().sum()

Out[158]: 0
```

Although, I have decided to perform transformations to handle outliers in the data, but it can be possible that during training the ML model, we don't get any improvement with and without transformation, but as a start I preferred to transform them and can decide things later based on the actual scenario.

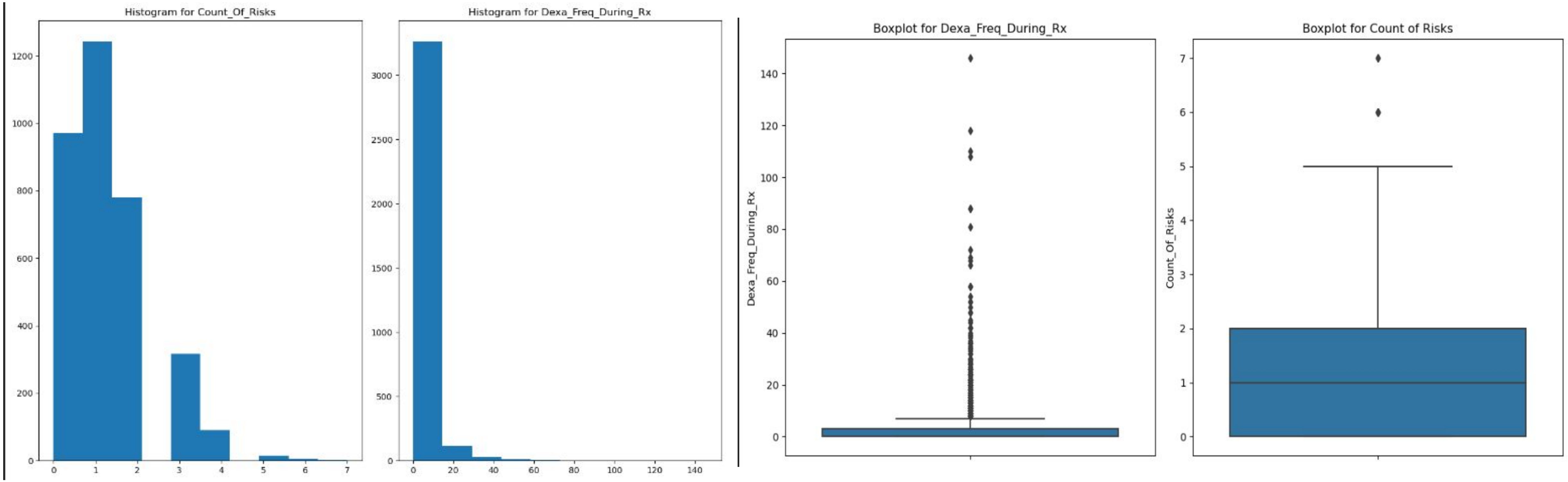
Individual Work by Tomisin

Data Glacier

Your Deep Learning Partner

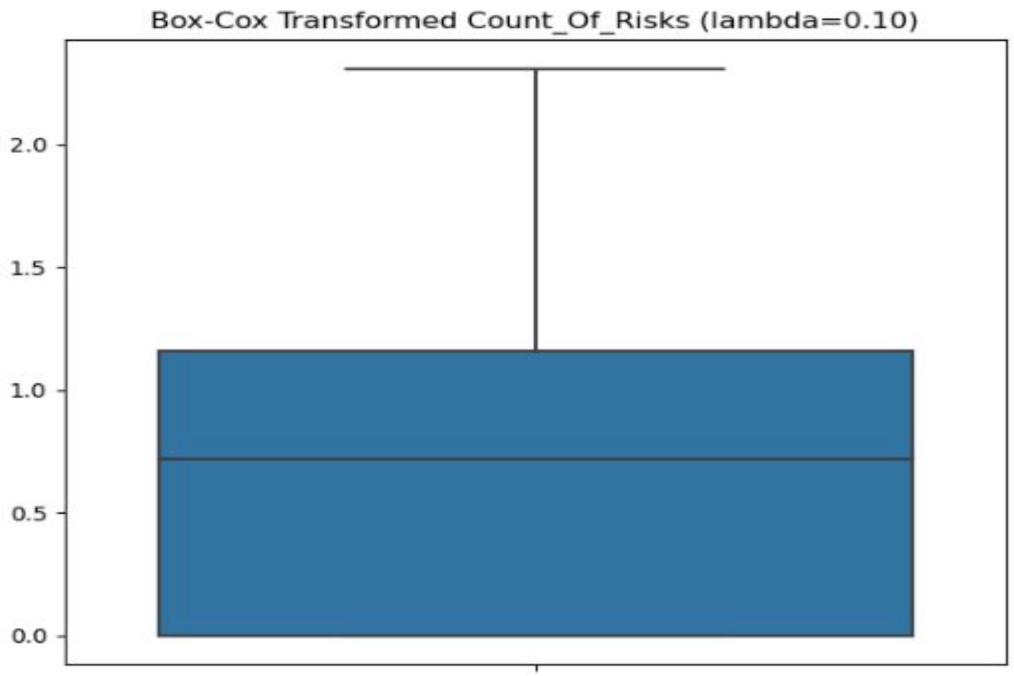
Data Preprocessing

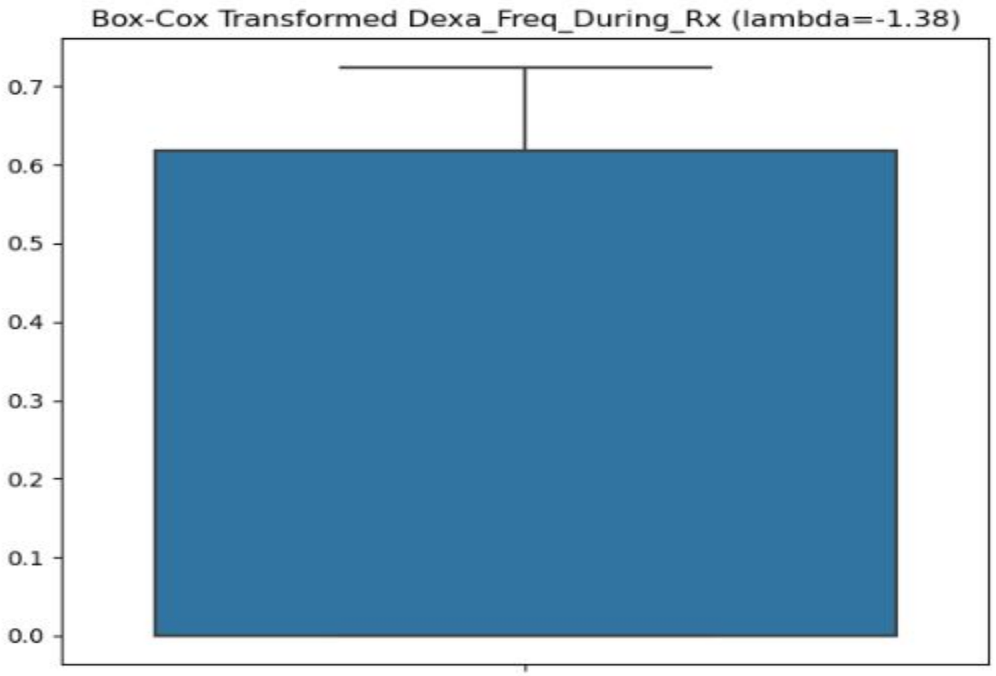
Count of Risks and Dexa Freq During Rx are both positively skewed and have outliers. Moreover, considering the size of the dataset, removing the outliers can impact the performance of the model, hence reducing variance seem to be a good option.



Data Preprocessing(cont.)

Boxplot after Box-Cox Transformation

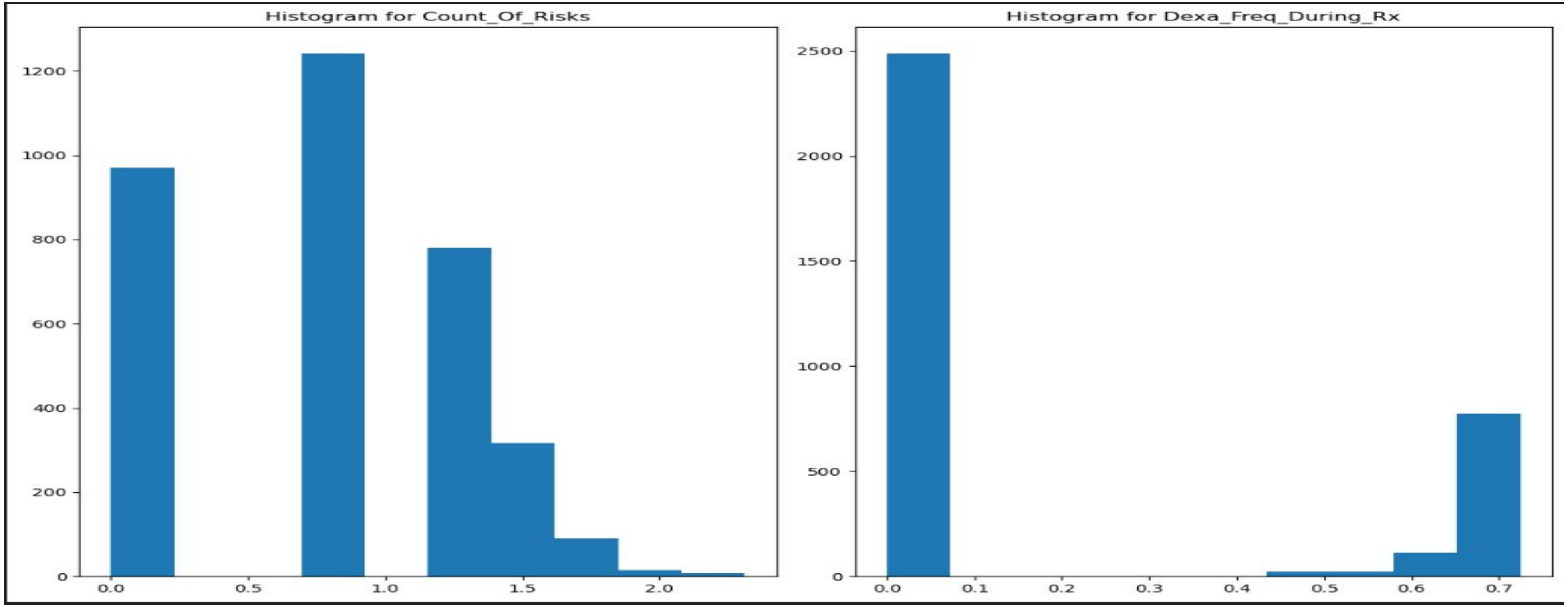




The outliers have been handled without losing any data

Data Preprocessing(cont.)

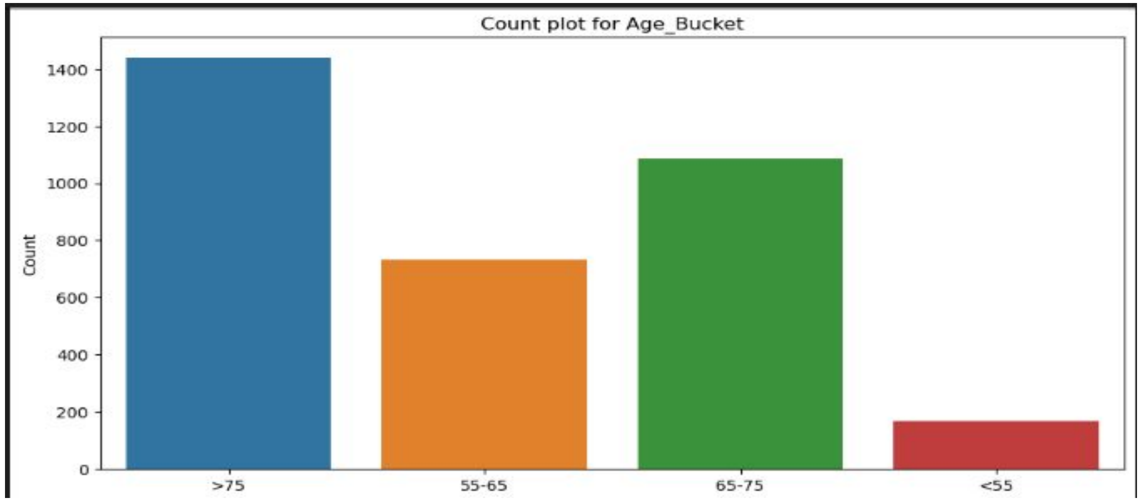
Histogram after Box-Cox Transformation



The spread of the data points have been increased

Data Preprocessing(cont.)

- Dropping ‘Ntm_Speciality’ and ‘Ntm_Speciality_Bucket’ considering ‘Ntm_flag’ might be sufficient.
- Reducing bins for Dexa_Freq_During_Rx to 6 to reduce outliers
- The dataset has more older generation, implying that old ones are more affected




```python
# Dropping redundant columns
df = df.drop(['Ptid', 'Ntm_Speciality', 'Ntm_Speciality_Bucket'], axis=1)

# Binning for 'Count_Of_Risks'
bins_count_of_risks = [0, 1, 2, 3, 4, float('inf')]
labels_count_of_risks = [0, 1, 2, 3, '>3']

df['Risk_Level'] = pd.cut(df['Count_Of_Risks'], bins=bins_count_of_risks, labels=labels_count_of_risks, right=False)
df = df.drop('Count_Of_Risks', axis=1)

# Binning for 'Dexa_Freq_During_Rx'
bins_dexa = [0, 6, 12, 18, 24, 30, float('inf')]
labels_dexa = [0, 6, 12, 18, 24, '>30']

df['Dexa_Freq_Level'] = pd.cut(df['Dexa_Freq_During_Rx'], bins=bins_dexa, labels=labels_dexa, right=False)
df = df.drop('Dexa_Freq_During_Rx', axis=1)
```

Thank You

Data Glacier

Your Deep Learning Partner