



**Data Glacier**

Your Deep Learning Partner

# Healthcare : Persistency of a Drug Final Project

## Virtual Internship

**Devin Chau, Ethan Dy, Rohan Khatri**



Healthcare: Persistency of a Drug			
Group Members	Devin Chau	Ethan Dy	Rohan Khatri
Email	chau.devin031602@gmail.com	ethan05dy@gmail.com	rohankhatri0507@gmail.com
Country	United States	United States	United States
Specialization	Data Science	Data Science	Data Science
Internship Batch	LISUM39	LISUM39	LISUM39
Date	14 January 2025	14 January 2025	14 January 2025

# Agenda

- Problem Statement
- Data Information
- Data Understanding
- Exploratory Data Analysis  
(EDA)
- Recommendations

# Problem Statement

## Context:

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification. With an objective to gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

## Problem Description:

We are building a predictive model that classifies patients into “persistent” or “non-persistent” categories based on factors like their demographics, medical history, physician characteristics, and treatment details. Factors like the patient level such as their age, risk factors, previous test results, or provider type allows for insights into why some patients continue therapy while others drop off. Thus understanding “persistence” levels. By analyzing these data points and finding patterns, the predictive model helps explain patient behavior and supports the creation of targeted interventions to improve adherence.

Ethan's [Github Repo Link](#)

# Data Information

<b>Total number of observations</b>	3424
<b>Total number of files</b>	1
<b>Total number of features</b>	69
<b>Base format of the file</b>	.csv
<b>Size of the data</b>	891 KB

# Data Information

Bucket	Variable	Variable Description
Unique Row Id	Patient ID	Unique ID of each patient
Target Variable	Persistency_Flag	Flag indicating if a patient was persistent or not
Demographics	Age	Age of the patient during their therapy
	Race	Race of the patient from the patient table
	Region	Region of the patient from the patient table
	Ethnicity	Ethnicity of the patient from the patient table
	Gender	Gender of the patient from the patient table
	IDN Indicator	Flag indicating patients mapped to IDN

# Data Information

Provider Attributes	NTM - Physician Specialty	Specialty of the HCP that prescribed the NTM Rx
	NTM - T-Score	T Score of the patient at the time of the NTM Rx (within 2 years prior from rxdate)
	Change in T Score	Change in Tscore before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown)
	NTM - Risk Segment	Risk Segment of the patient at the time of the NTM Rx (within 2 years days prior from rxdate)
	Change in Risk Segment	Change in Risk Segment before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown)
	NTM - Multiple Risk Factors	Flag indicating if patient falls under multiple risk category (having more than 1 risk) at the time of the NTM Rx (within 365 days prior from rxdate)



# Data Information

Clinical Factors	NTM - DEXA Scan Frequency	Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate)
	NTM - DEXA Scan Recency	Flag indicating the presence of DEXA Scan before the NTM Rx (within 2 years prior from rxdate or between their first Rx and Switched Rx; whichever is smaller and applicable)
	Dexa During Therapy	Flag indicating if the patient had a Dexa Scan during their first continuous therapy
	NTM - Fragility Fracture Recency	Flag indicating if the patient had a recent fragility fracture (within 365 days prior from rxdate)
	Fragility Fracture During Therapy	Flag indicating if the patient had fragility fracture during their first continuous therapy
	NTM - Glucocorticoid Recency	Flag indicating usage of Glucocorticoids ( $\geq 7.5$ mg strength) in the one year look-back from the first NTM Rx
	Glucocorticoid Usage During Therapy	Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy
	NTM - Injectable Experience	Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx
	NTM - Risk Factors	Risk Factors that the patient is falling into. For chronic Risk Factors complete lookback to be applied and for non-chronic Risk Factors, one year lookback from the date of first OP Rx

# Data Information

Disease/Treatment Factor	NTM - Comorbidity	Comorbidities are divided into two main categories - Acute and chronic, based on the ICD codes. For chronic disease we are taking complete look back from the first Rx date of NTM therapy and for acute diseases, time period before the NTM OP Rx with one year lookback has been applied
	NTM - Concomitancy	Concomitant drugs recorded prior to starting with a therapy(within 365 days prior from first rxdate)
	Adherence	Adherence for the therapies

# Data Understanding

#	Column	Non-Null	Count	Dtype
0	Ptid	3424	non-null	object
1	Persistency_Flag	3424	non-null	object
2	Gender	3424	non-null	object
3	Race	3424	non-null	object
4	Ethnicity	3424	non-null	object
5	Region	3424	non-null	object
6	Age_Bucket	3424	non-null	object
7	Ntm_Speciality	3424	non-null	object
8	Ntm_Specialist_Flag	3424	non-null	object
9	Ntm_Speciality_Bucket	3424	non-null	object
10	Gluko_Record_Prior_Ntm	3424	non-null	object
11	Gluko_Record_During_Rx	3424	non-null	object
12	Dexa_Freq_During_Rx	3424	non-null	int64
13	Dexa_During_Rx	3424	non-null	object
14	Frag_Frac_Prior_Ntm	3424	non-null	object
15	Frag_Frac_During_Rx	3424	non-null	object
16	Risk_Segment_Prior_Ntm	3424	non-null	object
17	Tscore_Bucket_Prior_Ntm	3424	non-null	object
18	Risk_Segment_During_Rx	3424	non-null	object
19	Tscore_Bucket_During_Rx	3424	non-null	object
20	Change_T_Score	3424	non-null	object
21	Change_Risk_Segment	3424	non-null	object
22	Adherent_Flag	3424	non-null	object
23	Idn_Indicator	3424	non-null	object
24	Injectable_Experience_During_Rx	3424	non-null	object
25	Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	3424	non-null	object
26	Comorb_Encounter_For_Immunization	3424	non-null	object
27	Comorb_Encntr_For_General_Exam_W_0_Complaint,_Susp_Or_Reprtd_Dx	3424	non-null	object
28	Comorb_Vitamin_D_Deficiency	3424	non-null	object
29	Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	3424	non-null	object
30	Comorb_Encntr_For_Oth_Sp_Exam_W_0_Complaint_Suspected_Or_Reprtd_Dx	3424	non-null	object
31	Comorb_Long_Term_Current_Drug_Therapy	3424	non-null	object
32	Comorb_Dorsalgia	3424	non-null	object
33	Comorb_Personal_History_Of_Other_Diseases_And_Conditions	3424	non-null	object
34	Comorb_Other_Disorders_Of_Bone_Density_And_Structure	3424	non-null	object
35	Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemia	3424	non-null	object
36	Comorb_Osteoporosis_without_current_pathological_fracture	3424	non-null	object
37	Comorb_Personal_history_of_malignant_neoplasm	3424	non-null	object
38	Comorb_Gastro_esophageal_reflux_disease	3424	non-null	object

- The dataset consists of 3424 rows and 69 columns
- Types of Variables
  - **Numeric (2 columns):**
    - Dexa\_Freq\_During\_Rx
    - Count\_Of\_Risks
  - **Categorical (67 columns):**
    - Examples: Persistency\_Flag, Gender, Ntm\_Speciality, etc
    - Many are binary flags (Y/N, etc.), while some have multiple categories (e.g., Ntm\_Speciality has 36)
- Multiple columns (such as Risk\_Segment\_During\_Rx, Change\_T\_Score, etc.) contain a large number of “Unknown” entries. This shows hidden missing data that could influence model training and interpretation

# Data Understanding

Missing Values Summary:

```
Column
Ptid
Concom_Cephalosporins
Risk_Osteogenesis Imperfecta
Risk_Type_1_Insulin_Dependent_Diabetes
Concom_Viral_Vaccines
...
Comorb_Other_Joint_Disorder_Not_Elsewhere_Class...
Comorb_Encntr_For_Oth_Sp_Exam_W_0_Complaint_Sus...
Comorb_Long_Term_Current_Drug_Therapy
Comorb_Dorsalgia
Count_Of_Risks
```

```
Missing Values \
Ptid 0
Concom_Cephalosporins 0
Risk_Osteogenesis Imperfecta 0
Risk_Type_1_Insulin_Dependent_Diabetes 0
Concom_Viral_Vaccines 0
...
Comorb_Other_Joint_Disorder_Not_Elsewhere_Class... 0
Comorb_Encntr_For_Oth_Sp_Exam_W_0_Complaint_Sus... 0
Comorb_Long_Term_Current_Drug_Therapy 0
Comorb_Dorsalgia 0
Count_Of_Risks 0
```

```
Percentage Missing
Ptid 0.0
Concom_Cephalosporins 0.0
Risk_Osteogenesis Imperfecta 0.0
Risk_Type_1_Insulin_Dependent_Diabetes 0.0
Concom_Viral_Vaccines 0.0
...
Comorb_Other_Joint_Disorder_Not_Elsewhere_Class... 0.0
Comorb_Encntr_For_Oth_Sp_Exam_W_0_Complaint_Sus... 0.0
Comorb_Long_Term_Current_Drug_Therapy 0.0
Comorb_Dorsalgia 0.0
Count_Of_Risks 0.0
```

[69 rows x 3 columns]

- Hidden “Unknown” Data

- Some columns use the string "Unknown" instead of NaN. Like, Risk\_Segment\_During\_Rx, Change\_T\_Score, and others have a lot of "Unknown" entries

- Outliers:

- Two numeric variables, Dexa\_Freq\_During\_Rx and Count\_Of\_Risks, have outliers:
  - Dexa\_Freq\_During\_Rx shows **460** outliers (based on the Interquartile Range method)
  - Count\_Of\_Risks shows **8** outliers (also IQR-based)

# Data Understanding

```
[ ] 1 scaler = MinMaxScaler()  
    2 encoded_data[numerical_cols] = scaler.fit_transform(encoded_data[numerical_cols])  
    3
```

- MinMaxScaler was used to normalize numerical columns, bringing all variables to a common scale and ensuring equal contribution to the model.
- **Feature Scaling:**
  - Scaled numerical features using MinMaxScaler to normalize their range, ensuring model fairness.

```
1 duplicates = df.duplicated(subset='PatientID').sum()
```

- Duplicate rows were detected based on the 'PatientID' column to prevent skewed analyses, and they were logged for further review.
- Columns were standardized to appropriate data types, reducing potential errors during analysis.

# Data Understanding

```
1 df = df[df['Dexa_Freq_During_Rx'] != 0]
```

- **Outlier Handling:**

- Removed outliers from the Dexa\_Freq\_During\_Rx column by excluding entries where its value was zero.

```
1 binary_columns = [col for col in df.columns if set(df[col].dropna().unique()) == {'N', 'Y'}]  
2  
3 for col in binary_columns:  
4     df[col] = df[col].replace({'N': 0, 'Y': 1}).astype(int)
```

- **Binary Encoding:**

- Converted categorical binary values ('N' and 'Y') into numerical representations (0 and 1), ensuring compatibility with classifier models.

# Data Understanding



```
1 # Using IQR
2 Q1 = df['Dexa_Freq_During_Rx'].quantile(0.25)
3 Q3 = df['Dexa_Freq_During_Rx'].quantile(0.75)
4 IQR = Q3 - Q1
5 lower_bound = Q1 - 1.5 * IQR
6 upper_bound = Q3 + 1.5 * IQR
7
8 df['Dexa_Freq_During_Rx_No_Outliers'] = np.where(
9     df['Dexa_Freq_During_Rx'] > upper_bound, upper_bound,
10     np.where(df['Dexa_Freq_During_Rx'] < lower_bound, lower_bound, df['Dexa_Freq_During_Rx']))
11 )
12 df['Dexa_Freq_During_Rx_No_Outliers'].describe()
```

- **Outlier Handling:**
  - Applied the IQR method to cap extreme values within acceptable ranges, addressing potential data skew.

# Data Understanding

```
1 risk_bins = [-1, 0, 2, 5, float('inf')]  
2 risk_labels = ['None', 'Low', 'Moderate', 'High']  
3 encoded_data['Risk_Level'] = pd.cut(encoded_data['Count_Of_Risks'].astype(int),  
4                                     bins=risk_bins,  
5                                     labels=risk_labels)  
6
```

- **Risk Level Encoding:**
  - Categorized patient risk levels into bins such as 'None', 'Low', 'Moderate', and 'High' based on Count\_Of\_Risks values, enabling targeted risk analysis.

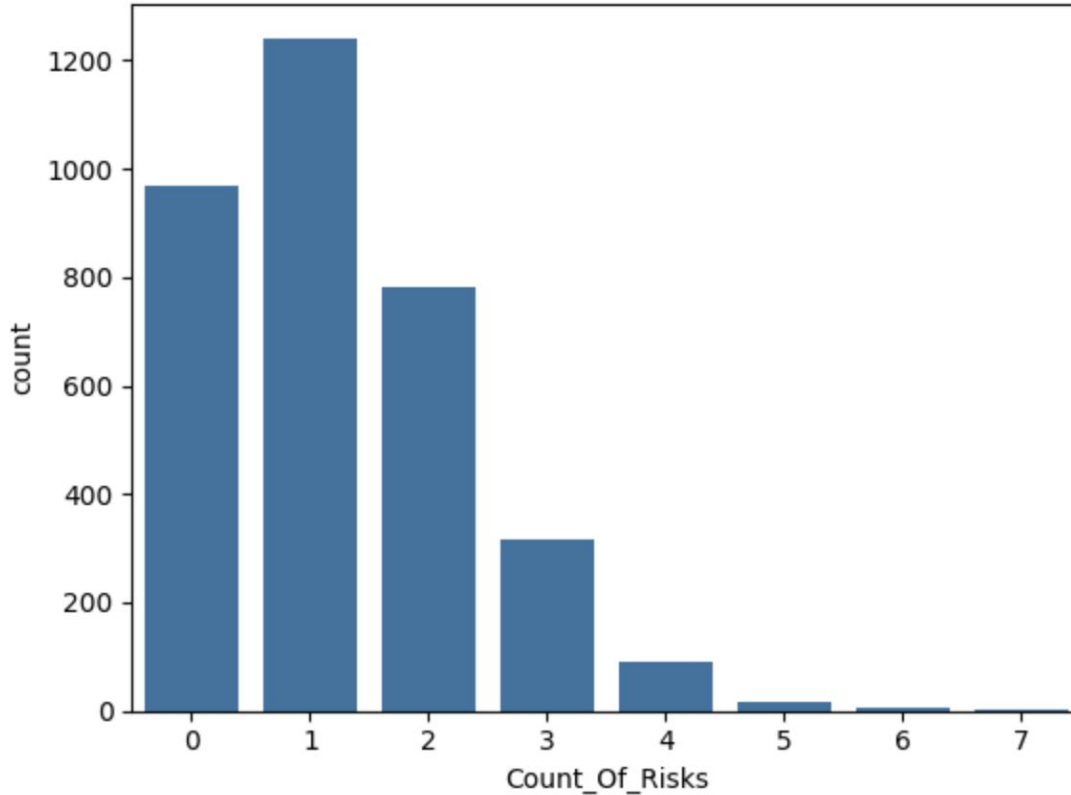


# Exploratory Data Analysis (EDA)

During the exploratory data analysis (EDA), we performed the following steps:

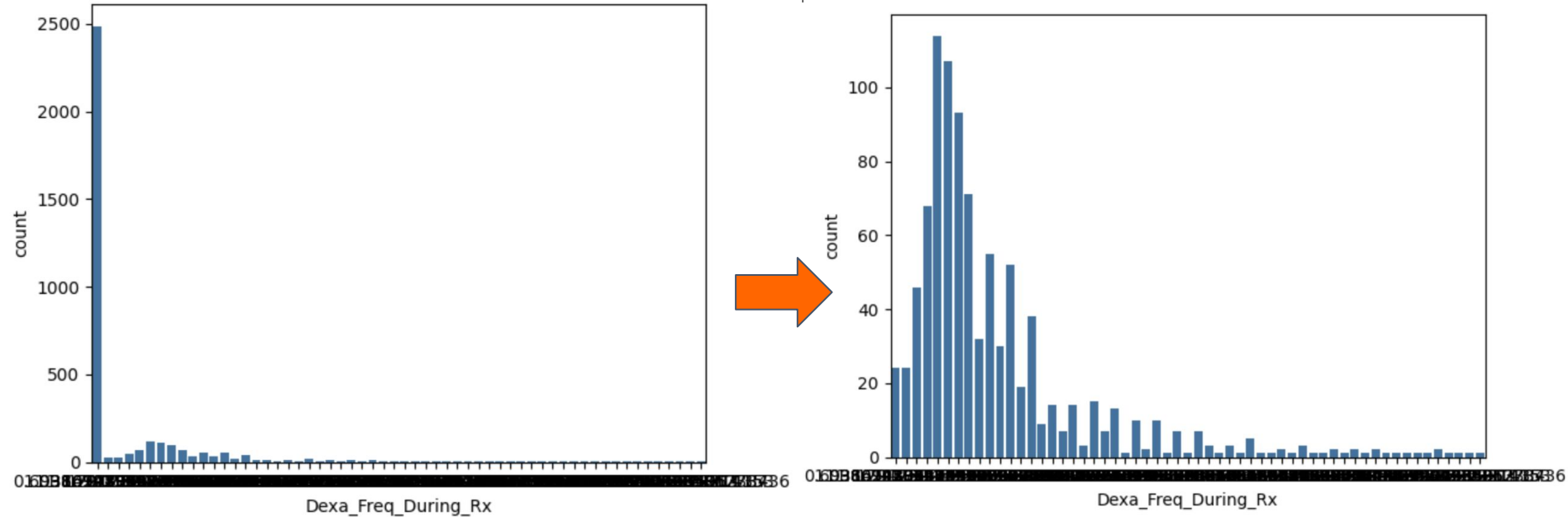
- **Data Visualization:** Visualized distributions of numerical features such as Dexa\_Freq\_During\_Rx and Count\_Of\_Risks.
  - Histograms showed high skewness in Dexa\_Freq\_During\_Rx, leading to the decision to apply log transformation for normalization
- **Correlation Analysis:** Calculated correlations between numerical variables. Count\_Of\_Risks showed a moderate correlation with the target variable Persistency\_Flag
- **Categorical Feature Inspection:** Explored distributions for categorical variables like Ntm\_Speciality and Gender. Identified the imbalanced classes in Ntm\_Speciality which could maybe lead to overfitting (?)
- **Outlier Detection:** Used the Interquartile Range (IQR) method to detect outliers. Dexa\_Freq\_During\_Rx had 460 extreme values,
  - while Count\_Of\_Risks had 8 outliers (addressed using capping methods)
- **Handling Missing Values:** Discovered multiple columns with the value "Unknown" instead of NaNs, indicating hidden missing data. Handled as separate categories

# Exploratory Data Analysis (EDA)



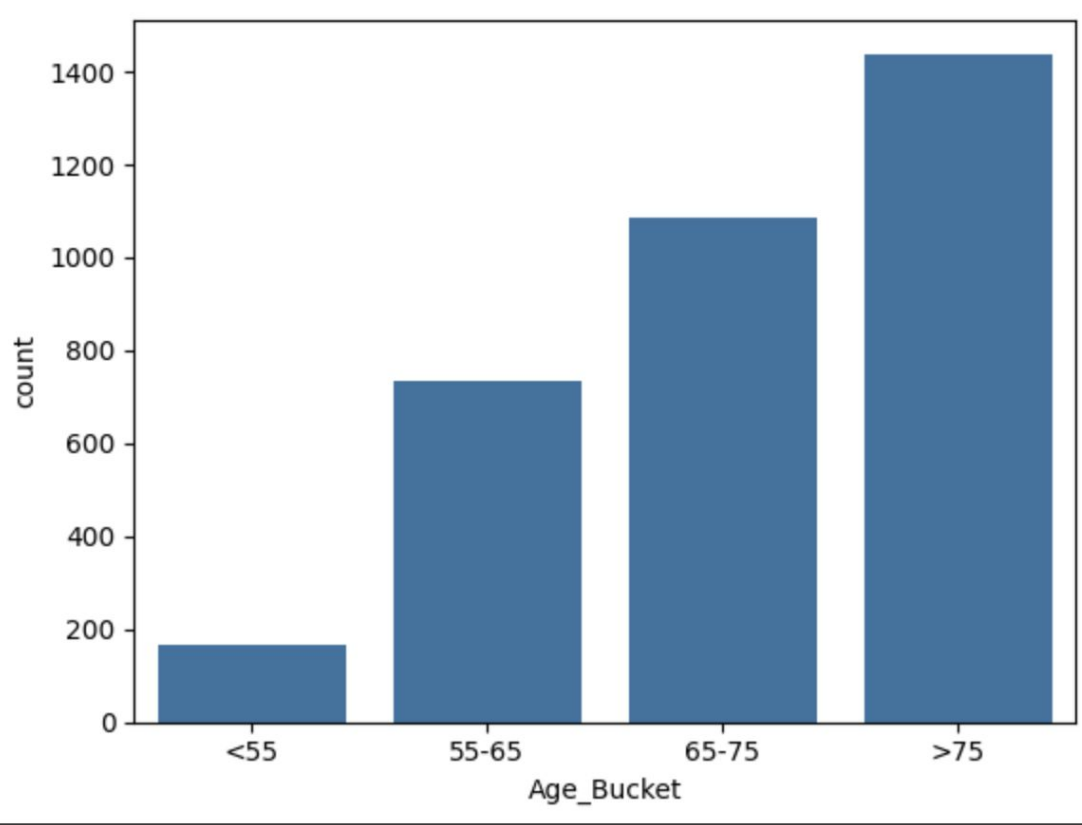
- As we can see, most people tend to have 0-2 counts of risk in total, which is good to see. It is better to see less counts of risks in comparison to the higher numbers.

# Exploratory Data Analysis (EDA)



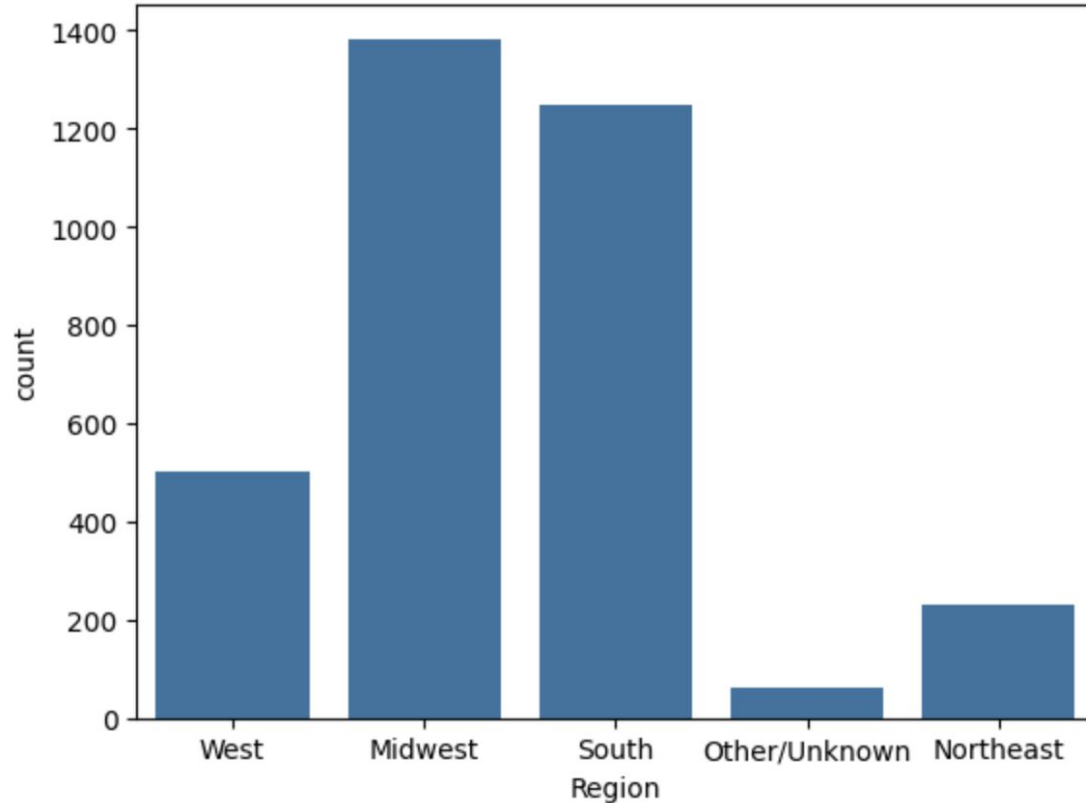
- This is what the distribution looks like without the outlier of 0

# Exploratory Data Analysis (EDA)



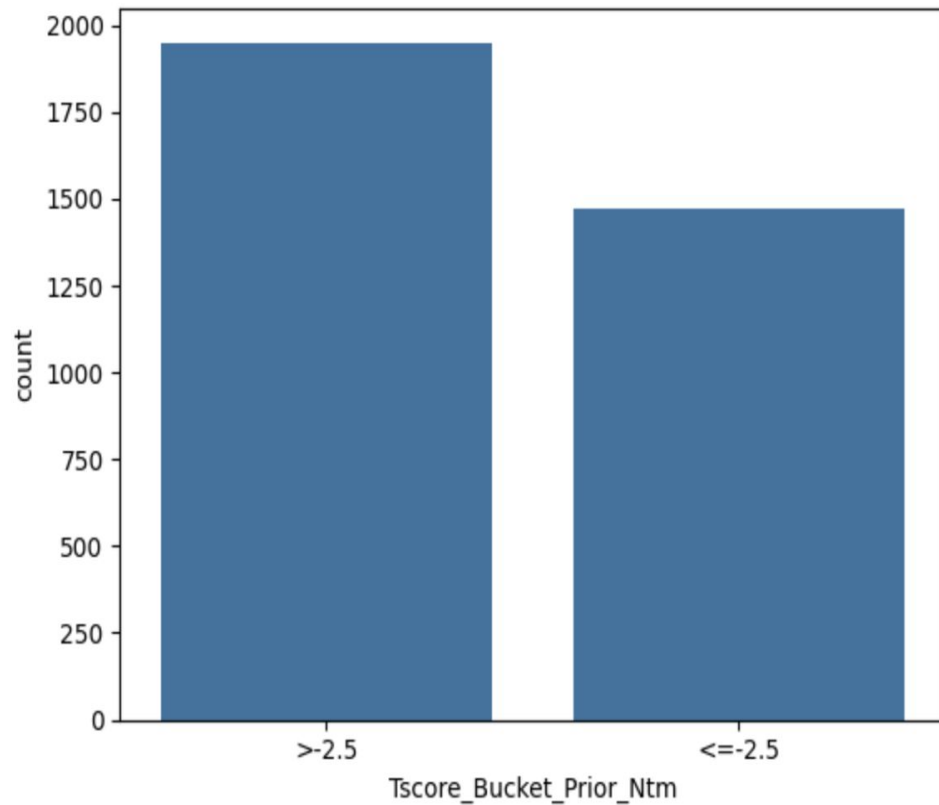
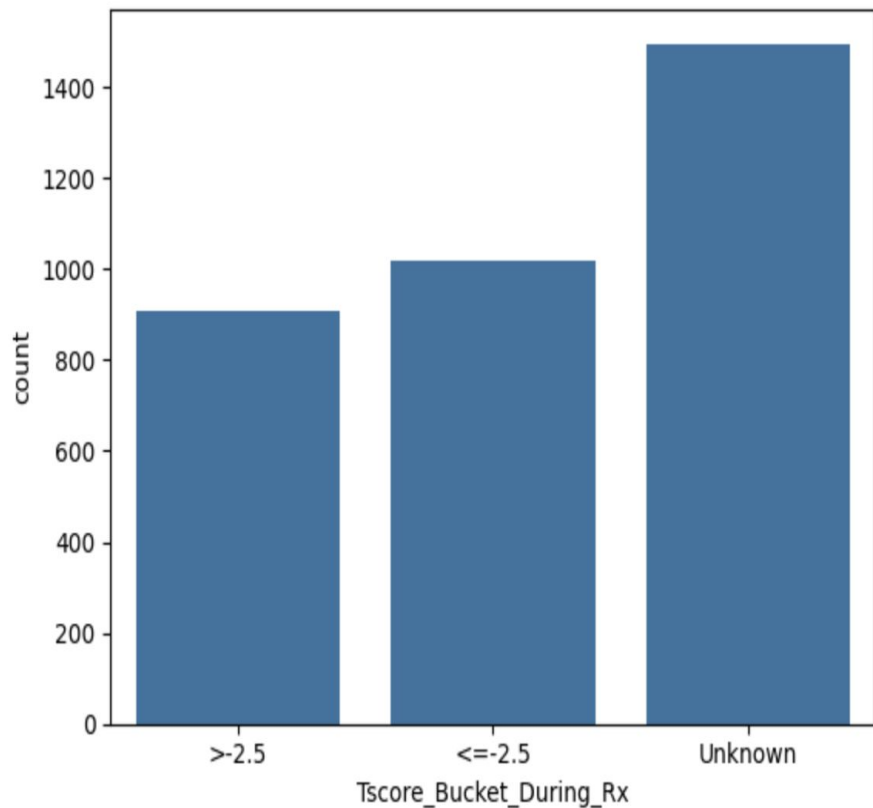
- The largest proportion of patients reported in this dataset belongs to the older age group, specifically those aged >75.

# Exploratory Data Analysis (EDA)

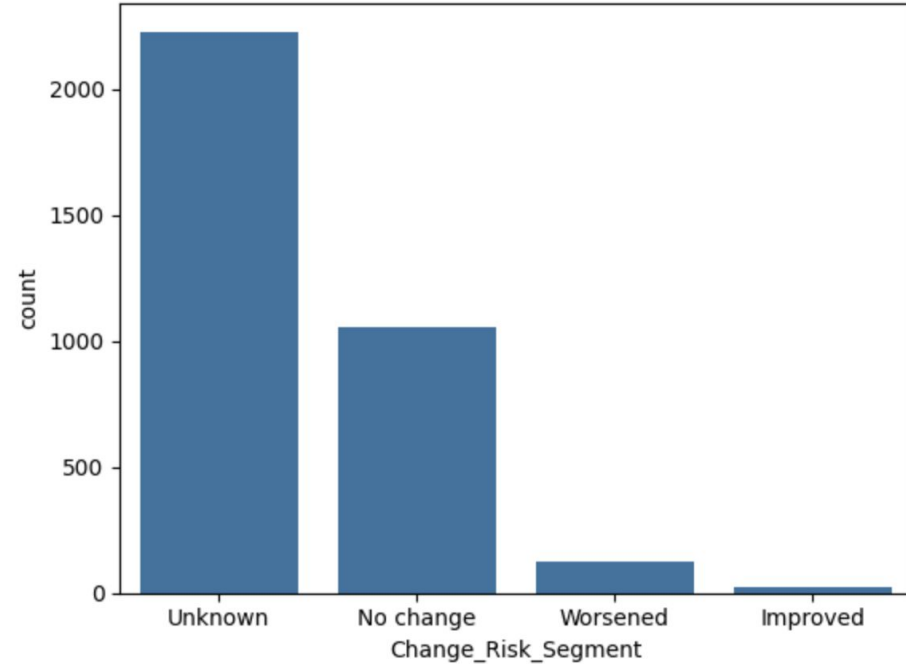
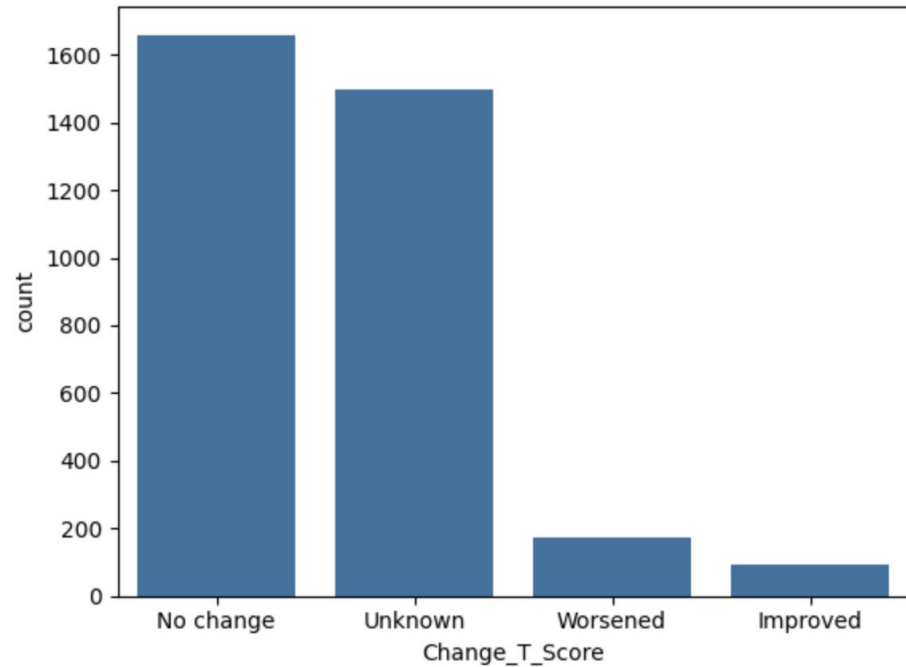


- Northeast and West seem to be severely underreported as compared to Midwest and South Region

# Exploratory Data Analysis (EDA)

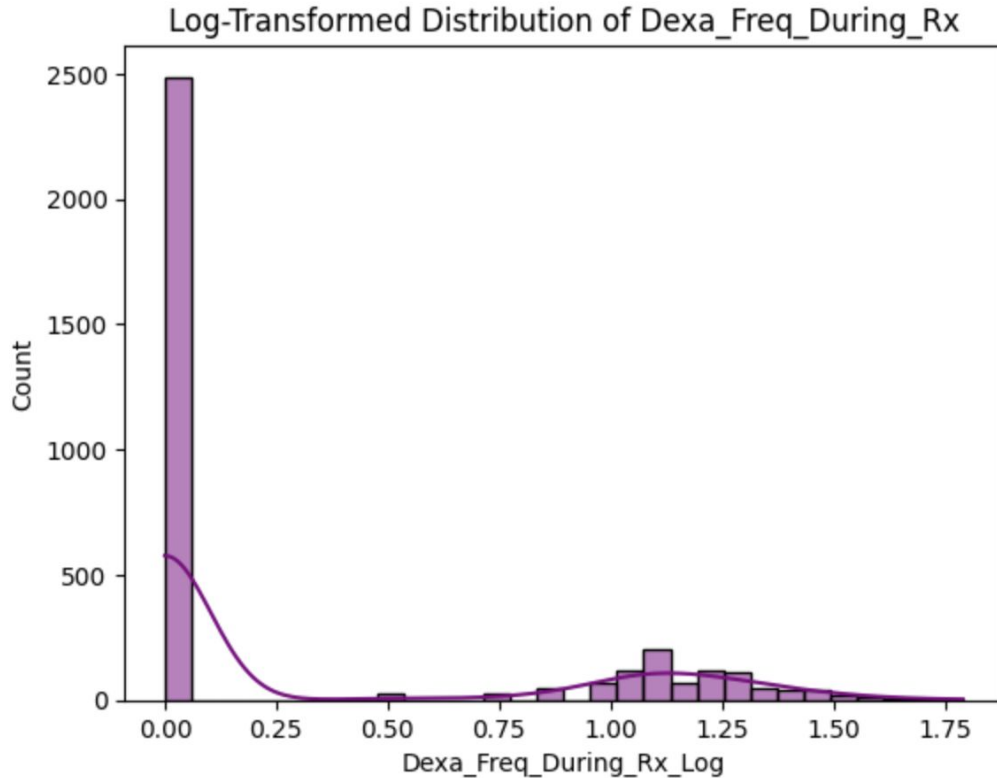


# Exploratory Data Analysis (EDA)



- We can see that typically there is no change. Worsened more than improved in terms of change risk segment.

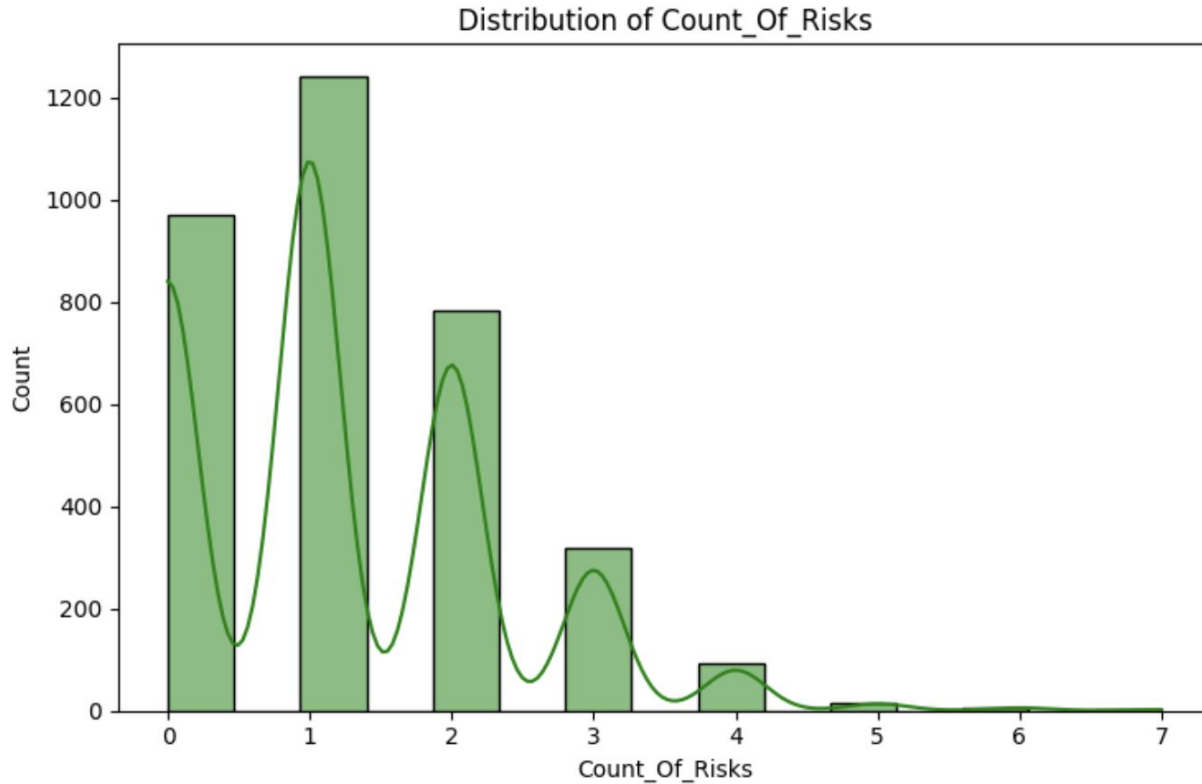
# Exploratory Data Analysis (EDA)



- Log transformation to Dexa\_Freq\_During\_Rx for normalization

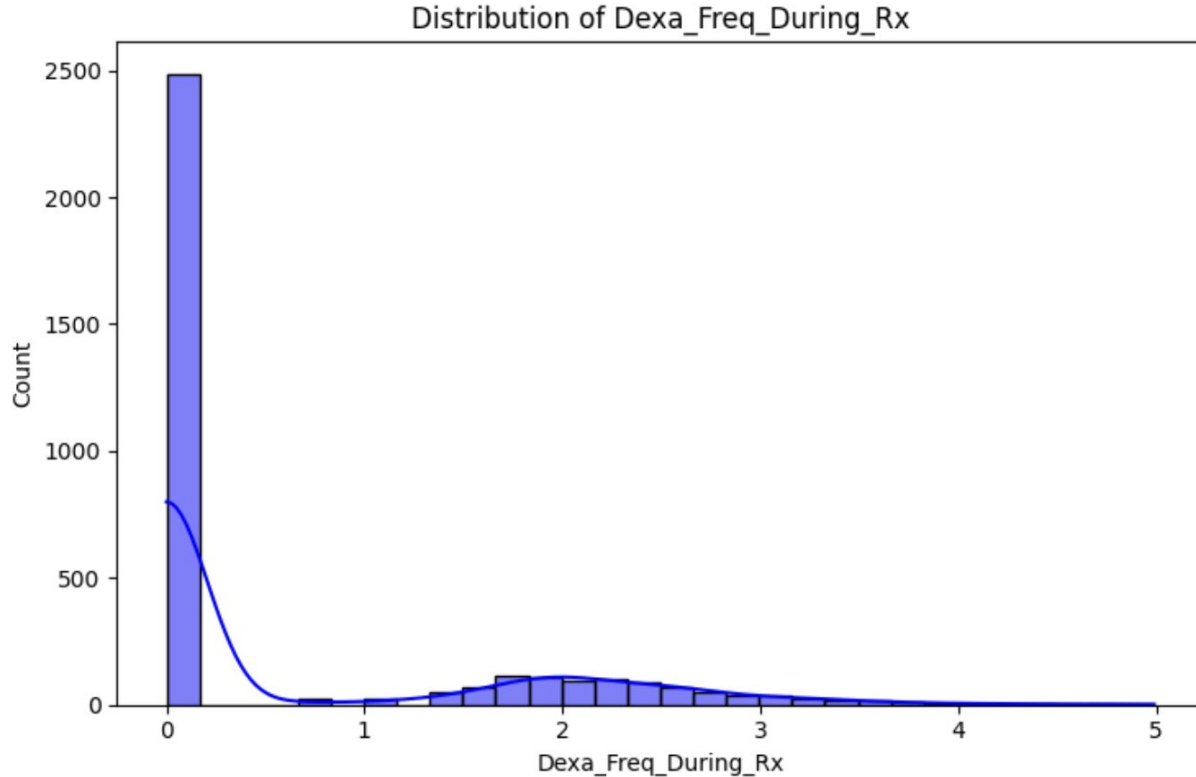


# Exploratory Data Analysis (EDA)



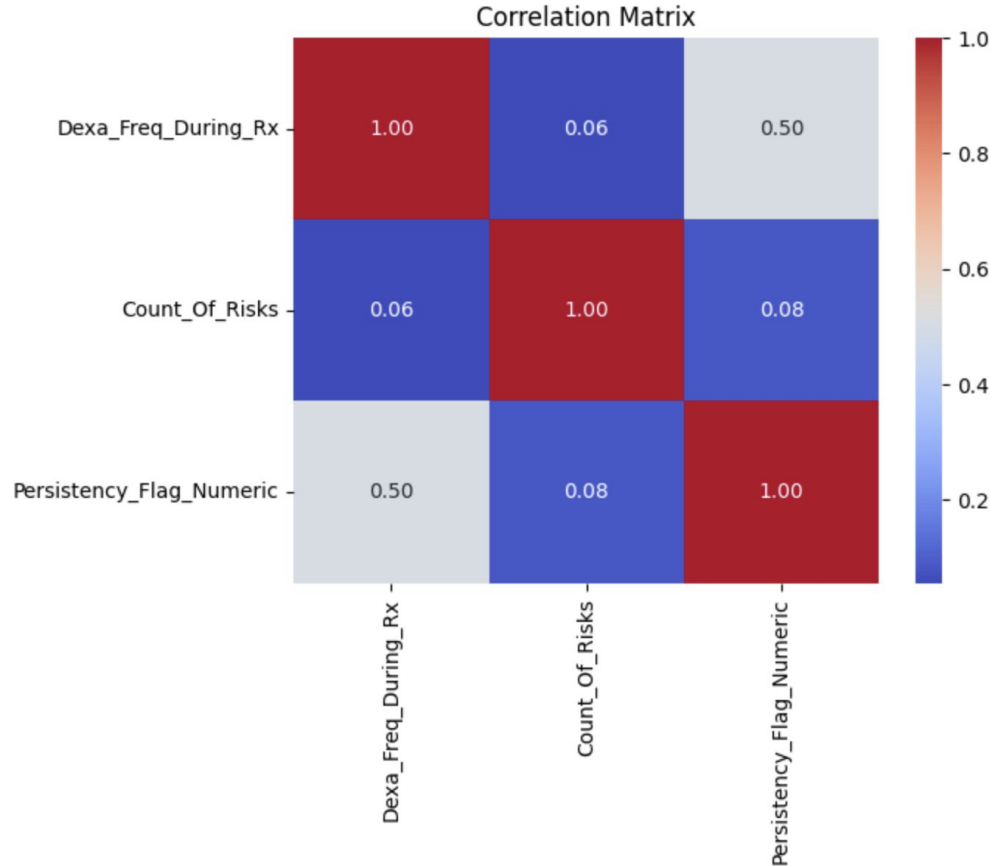
- Histogram for Count\_Of\_Risks

# Exploratory Data Analysis (EDA)



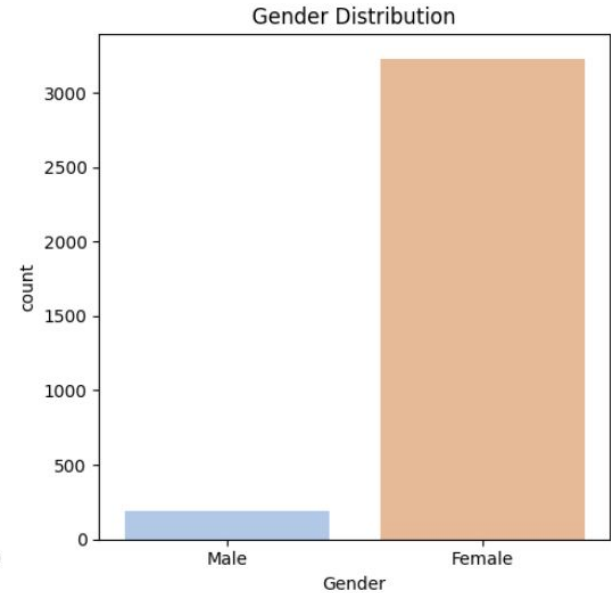
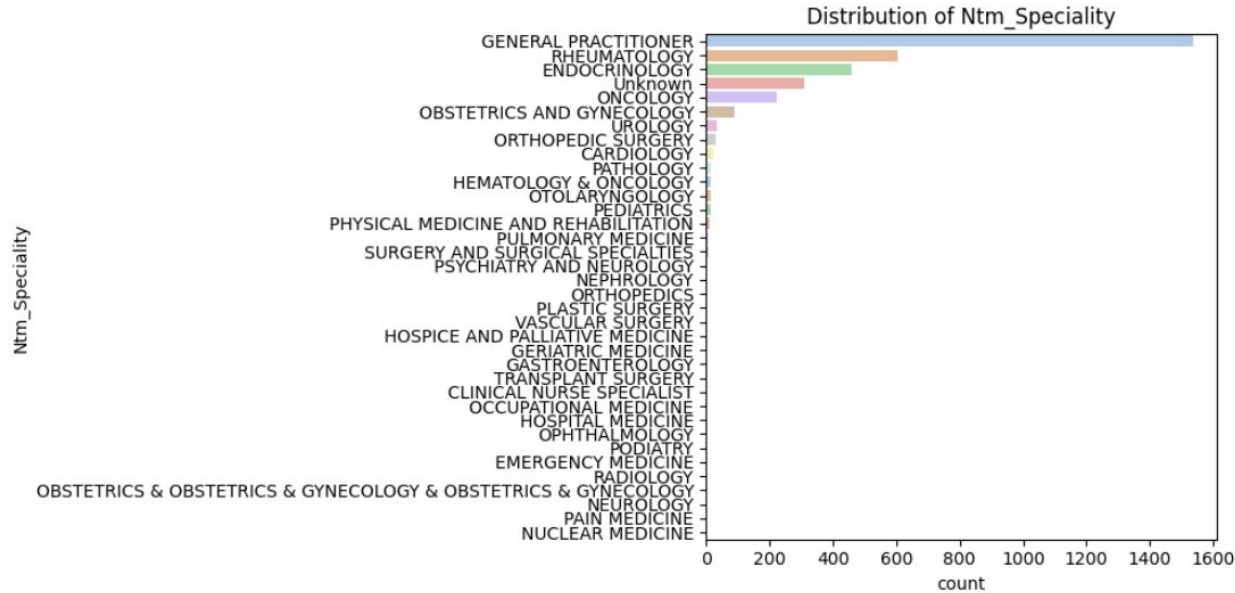
- Histogram for normalized Dexa\_Freq\_During\_Rx

# Exploratory Data Analysis (EDA)



- The variables are mostly uncorrelated except for a moderate relationship between Dexa\_Freq\_During\_Rx and Persistency\_Flag\_Numeric.
- Higher Dexa scan frequency might be slightly associated with better persistence in treatment adherence.
- Count of risks does not show a meaningful correlation with either of the other variables.

# Exploratory Data Analysis (EDA)



- The data suggests that most observations come from general practitioners and predominantly female participants.
- The gender imbalance might influence study outcomes if gender plays a role in the analysis being conducted.

# Exploratory Data Analysis (EDA)

## Final Recommendations

Based on the analysis and identified data issues, we recommend:

- **Handling Missing Values:** Maintain “Unknown” entries as a separate category to preserve data integrity and avoid loss of potentially valuable patterns
- **Normalization:** Continue using MinMaxScaler for numerical features like DEXA\_Freq\_During\_Rx and Count\_Of\_Risks to ensure consistent scaling across variables
- **Outlier Handling:** Use the IQR method to cap extreme values for DEXA\_Freq\_During\_Rx and Count\_Of\_Risks
- **Categorical Encoding:** Apply one-hot encoding for categorical variables with multiple categories. Using binary encoding for columns with simple Y/N values
- **Feature Engineering:** Consider grouping rare categories in columns like Ntm\_Speciality to avoid overfitting due to high cardinality
- **Data Consistency:** Ensure proper standardization of data types and consistency across the entire dataset before model training

# Exploratory Data Analysis (EDA)

## Recommended models for this datasets:

- **Logistic Regression:**
  - a. For binary classification.
- **Decision Trees:**
  - a. Easy to interpret and handle categorical variables directly.
- **Random Forest:**
  - Robust and simple, handling mixed data types well.
    - i. We will probably look into random forest more so than decision trees

# Thank You