Week 9_ WanguNdungu

August 2, 2022

1 Drug Persistency and Medical Adherence

Team Name: Feature Transformers Team members: - Name: Wangu Ndungu - Email: nwangu349@gmail.com - Country: Kenya - College/Company: Kenyatta University - Specialization: Data Science

- Name: Nikola Andrejić
- Email: nikola.ing.nl@gmail.com
- Country: Serbia
- College/Company: University of Niš
- Specialization: Data Science

1.1 PROBLEM STATEMENT

According to the World Health Organisation, only 50-70% of patients adhere properly to prescribed drugs during therapy. This is especially true among those with long term medication. This worrying statistic is caused by various factors, for example: patient's condition or disease, their socio-economic status, confusion by the schedule, forgetting, discontinuing because they feel better, just to name a few. Medical non-adherence can lead to devastating consequences on one's health, especially those with chronic illnesses. The purpose of this project is to study trends among patients in a sample and build a model that'll classify a new patient as Persistent or Non-Persistent. This project will givemedical practitioners(especially pharmaceuticals) insight on which patients might require more rigorous follow-ups to ensure they will adhere to their prescriptions.

1.1.1 Importing the required libraries

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import skew, stats
%matplotlib inline
```

1.1.2 Importing the data

[2]:
```
df = pd.read_excel('C:/Users/user/Drug percistency/Healthcare_dataset.xlsx',␣
↪sheet_name='Dataset')
df.head(5)
```

[2]:

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region
0	P1	Persistent	Male	Caucasian	Not Hispanic	West
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest

	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag
0	>75	GENERAL PRACTITIONER	Others
1	55-65	GENERAL PRACTITIONER	Others
2	65-75	GENERAL PRACTITIONER	Others
3	>75	GENERAL PRACTITIONER	Others
4	>75	GENERAL PRACTITIONER	Others

	Ntm_Speciality_Bucket	…	Risk_Family_History_Of_Osteoporosis
0	OB/GYN/Others/PCP/Unknown	…	N
1	OB/GYN/Others/PCP/Unknown	…	N
2	OB/GYN/Others/PCP/Unknown	…	N
3	OB/GYN/Others/PCP/Unknown	…	N
4	OB/GYN/Others/PCP/Unknown	…	N

	Risk_Low_Calcium_Intake	Risk_Vitamin_D_Insufficiency
0	N	N
1	N	N
2	Y	N
3	N	N
4	N	N

	Risk_Poor_Health_Frailty	Risk_Excessive_Thinness
0	N	N
1	N	N
2	N	N
3	N	N
4	N	N

	Risk_Hysterectomy_Oophorectomy	Risk_Estrogen_Deficiency	Risk_Immobilization
0	N	N	N
1	N	N	N
2	N	N	N
3	N	N	N
4	N	N	N

	Risk_Recurring_Falls	Count_Of_Risks
0	N	0
1	N	0
2	N	2
3	N	1

```
4                    N              1

[5 rows x 69 columns]
```

1.1.3 Data attributes

[3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3424 entries, 0 to 3423
Data columns (total 69 columns):
 #   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  Gluco_Record_Prior_Ntm                                    3424
non-null   object
 11  Gluco_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_O_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_O_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_lipidemias  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
 39  Concom_Cholesterol_And_Triglyceride_Regulating_Preparations  3424
non-null   object
 40  Concom_Narcotics                                           3424
```

```
non-null   object
 41  Concom_Systemic_Corticosteroids_Plain                 3424
non-null   object
 42  Concom_Anti_Depressants_And_Mood_Stabilisers          3424
non-null   object
 43  Concom_Fluoroquinolones                               3424
non-null   object
 44  Concom_Cephalosporins                                 3424
non-null   object
 45  Concom_Macrolides_And_Similar_Types                   3424
non-null   object
 46  Concom_Broad_Spectrum_Penicillins                     3424
non-null   object
 47  Concom_Anaesthetics_General                           3424
non-null   object
 48  Concom_Viral_Vaccines                                 3424
non-null   object
 49  Risk_Type_1_Insulin_Dependent_Diabetes                3424
non-null   object
 50  Risk_Osteogenesis_Imperfecta                          3424
non-null   object
 51  Risk_Rheumatoid_Arthritis                             3424
non-null   object
 52  Risk_Untreated_Chronic_Hyperthyroidism                3424
non-null   object
 53  Risk_Untreated_Chronic_Hypogonadism                   3424
non-null   object
 54  Risk_Untreated_Early_Menopause                        3424
non-null   object
 55  Risk_Patient_Parent_Fractured_Their_Hip               3424
non-null   object
 56  Risk_Smoking_Tobacco                                  3424
non-null   object
 57  Risk_Chronic_Malnutrition_Or_Malabsorption            3424
non-null   object
 58  Risk_Chronic_Liver_Disease                            3424
non-null   object
 59  Risk_Family_History_Of_Osteoporosis                   3424
non-null   object
 60  Risk_Low_Calcium_Intake                               3424
non-null   object
 61  Risk_Vitamin_D_Insufficiency                          3424
non-null   object
 62  Risk_Poor_Health_Frailty                              3424
non-null   object
 63  Risk_Excessive_Thinness                               3424
non-null   object
 64  Risk_Hysterectomy_Oophorectomy                        3424
```

```
non-null   object
 65  Risk_Estrogen_Deficiency                                    3424
non-null   object
 66  Risk_Immobilization                                         3424
non-null   object
 67  Risk_Recurring_Falls                                        3424
non-null   object
 68  Count_Of_Risks                                              3424
non-null   int64
dtypes: int64(2), object(67)
memory usage: 1.8+ MB
```

This dataset has a total of 68 variables, Percistency_Flag being our target variable. 67 of these attributes are categorical and 2 of the are continuous.

1.1.4 Checking for missing data

[4]:
```
missing_data = df.isnull()
missing_data.head(5)
```

[4]:

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	\
0	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	

	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	…	\
0	False	False	False	…	
1	False	False	False	…	
2	False	False	False	…	
3	False	False	False	…	
4	False	False	False	…	

	Risk_Family_History_Of_Osteoporosis	Risk_Low_Calcium_Intake	\
0	False	False	
1	False	False	
2	False	False	
3	False	False	
4	False	False	

	Risk_Vitamin_D_Insufficiency	Risk_Poor_Health_Frailty	\
0	False	False	
1	False	False	
2	False	False	
3	False	False	
4	False	False	

```
   Risk_Excessive_Thinness  Risk_Hysterectomy_Oophorectomy  \
0                    False                           False
1                    False                           False
2                    False                           False
3                    False                           False
4                    False                           False

   Risk_Estrogen_Deficiency  Risk_Immobilization  Risk_Recurring_Falls  \
0                     False                False                 False
1                     False                False                 False
2                     False                False                 False
3                     False                False                 False
4                     False                False                 False

   Count_Of_Risks
0           False
1           False
2           False
3           False
4           False

[5 rows x 69 columns]
```

[5]:
```
for column in missing_data.columns.values.tolist():
    print(column)
    print (missing_data[column].value_counts())
    print("")
```

```
Ptid
False    3424
Name: Ptid, dtype: int64

Persistency_Flag
False    3424
Name: Persistency_Flag, dtype: int64

Gender
False    3424
Name: Gender, dtype: int64

Race
False    3424
Name: Race, dtype: int64

Ethnicity
False    3424
Name: Ethnicity, dtype: int64
```

```
Region
False    3424
Name: Region, dtype: int64

Age_Bucket
False    3424
Name: Age_Bucket, dtype: int64

Ntm_Speciality
False    3424
Name: Ntm_Speciality, dtype: int64

Ntm_Specialist_Flag
False    3424
Name: Ntm_Specialist_Flag, dtype: int64

Ntm_Speciality_Bucket
False    3424
Name: Ntm_Speciality_Bucket, dtype: int64

Gluco_Record_Prior_Ntm
False    3424
Name: Gluco_Record_Prior_Ntm, dtype: int64

Gluco_Record_During_Rx
False    3424
Name: Gluco_Record_During_Rx, dtype: int64

Dexa_Freq_During_Rx
False    3424
Name: Dexa_Freq_During_Rx, dtype: int64

Dexa_During_Rx
False    3424
Name: Dexa_During_Rx, dtype: int64

Frag_Frac_Prior_Ntm
False    3424
Name: Frag_Frac_Prior_Ntm, dtype: int64

Frag_Frac_During_Rx
False    3424
Name: Frag_Frac_During_Rx, dtype: int64

Risk_Segment_Prior_Ntm
False    3424
Name: Risk_Segment_Prior_Ntm, dtype: int64
```

```
Tscore_Bucket_Prior_Ntm
False    3424
Name: Tscore_Bucket_Prior_Ntm, dtype: int64

Risk_Segment_During_Rx
False    3424
Name: Risk_Segment_During_Rx, dtype: int64

Tscore_Bucket_During_Rx
False    3424
Name: Tscore_Bucket_During_Rx, dtype: int64

Change_T_Score
False    3424
Name: Change_T_Score, dtype: int64

Change_Risk_Segment
False    3424
Name: Change_Risk_Segment, dtype: int64

Adherent_Flag
False    3424
Name: Adherent_Flag, dtype: int64

Idn_Indicator
False    3424
Name: Idn_Indicator, dtype: int64

Injectable_Experience_During_Rx
False    3424
Name: Injectable_Experience_During_Rx, dtype: int64

Comorb_Encounter_For_Screening_For_Malignant_Neoplasms
False    3424
Name: Comorb_Encounter_For_Screening_For_Malignant_Neoplasms, dtype: int64

Comorb_Encounter_For_Immunization
False    3424
Name: Comorb_Encounter_For_Immunization, dtype: int64

Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx
False    3424
Name: Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx, dtype:
int64

Comorb_Vitamin_D_Deficiency
False    3424
Name: Comorb_Vitamin_D_Deficiency, dtype: int64
```

```
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified
False    3424
Name: Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified, dtype: int64

Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx
False    3424
Name: Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx, dtype:
int64

Comorb_Long_Term_Current_Drug_Therapy
False    3424
Name: Comorb_Long_Term_Current_Drug_Therapy, dtype: int64

Comorb_Dorsalgia
False    3424
Name: Comorb_Dorsalgia, dtype: int64

Comorb_Personal_History_Of_Other_Diseases_And_Conditions
False    3424
Name: Comorb_Personal_History_Of_Other_Diseases_And_Conditions, dtype: int64

Comorb_Other_Disorders_Of_Bone_Density_And_Structure
False    3424
Name: Comorb_Other_Disorders_Of_Bone_Density_And_Structure, dtype: int64

Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias
False    3424
Name: Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias, dtype:
int64

Comorb_Osteoporosis_without_current_pathological_fracture
False    3424
Name: Comorb_Osteoporosis_without_current_pathological_fracture, dtype: int64

Comorb_Personal_history_of_malignant_neoplasm
False    3424
Name: Comorb_Personal_history_of_malignant_neoplasm, dtype: int64

Comorb_Gastro_esophageal_reflux_disease
False    3424
Name: Comorb_Gastro_esophageal_reflux_disease, dtype: int64

Concom_Cholesterol_And_Triglyceride_Regulating_Preparations
False    3424
Name: Concom_Cholesterol_And_Triglyceride_Regulating_Preparations, dtype: int64

Concom_Narcotics
```

```
False    3424
Name: Concom_Narcotics, dtype: int64

Concom_Systemic_Corticosteroids_Plain
False    3424
Name: Concom_Systemic_Corticosteroids_Plain, dtype: int64

Concom_Anti_Depressants_And_Mood_Stabilisers
False    3424
Name: Concom_Anti_Depressants_And_Mood_Stabilisers, dtype: int64

Concom_Fluoroquinolones
False    3424
Name: Concom_Fluoroquinolones, dtype: int64

Concom_Cephalosporins
False    3424
Name: Concom_Cephalosporins, dtype: int64

Concom_Macrolides_And_Similar_Types
False    3424
Name: Concom_Macrolides_And_Similar_Types, dtype: int64

Concom_Broad_Spectrum_Penicillins
False    3424
Name: Concom_Broad_Spectrum_Penicillins, dtype: int64

Concom_Anaesthetics_General
False    3424
Name: Concom_Anaesthetics_General, dtype: int64

Concom_Viral_Vaccines
False    3424
Name: Concom_Viral_Vaccines, dtype: int64

Risk_Type_1_Insulin_Dependent_Diabetes
False    3424
Name: Risk_Type_1_Insulin_Dependent_Diabetes, dtype: int64

Risk_Osteogenesis_Imperfecta
False    3424
Name: Risk_Osteogenesis_Imperfecta, dtype: int64

Risk_Rheumatoid_Arthritis
False    3424
Name: Risk_Rheumatoid_Arthritis, dtype: int64

Risk_Untreated_Chronic_Hyperthyroidism
```

```
False    3424
Name: Risk_Untreated_Chronic_Hyperthyroidism, dtype: int64

Risk_Untreated_Chronic_Hypogonadism
False    3424
Name: Risk_Untreated_Chronic_Hypogonadism, dtype: int64

Risk_Untreated_Early_Menopause
False    3424
Name: Risk_Untreated_Early_Menopause, dtype: int64

Risk_Patient_Parent_Fractured_Their_Hip
False    3424
Name: Risk_Patient_Parent_Fractured_Their_Hip, dtype: int64

Risk_Smoking_Tobacco
False    3424
Name: Risk_Smoking_Tobacco, dtype: int64

Risk_Chronic_Malnutrition_Or_Malabsorption
False    3424
Name: Risk_Chronic_Malnutrition_Or_Malabsorption, dtype: int64

Risk_Chronic_Liver_Disease
False    3424
Name: Risk_Chronic_Liver_Disease, dtype: int64

Risk_Family_History_Of_Osteoporosis
False    3424
Name: Risk_Family_History_Of_Osteoporosis, dtype: int64

Risk_Low_Calcium_Intake
False    3424
Name: Risk_Low_Calcium_Intake, dtype: int64

Risk_Vitamin_D_Insufficiency
False    3424
Name: Risk_Vitamin_D_Insufficiency, dtype: int64

Risk_Poor_Health_Frailty
False    3424
Name: Risk_Poor_Health_Frailty, dtype: int64

Risk_Excessive_Thinness
False    3424
Name: Risk_Excessive_Thinness, dtype: int64

Risk_Hysterectomy_Oophorectomy
```

```
False    3424
Name: Risk_Hysterectomy_Oophorectomy, dtype: int64

Risk_Estrogen_Deficiency
False    3424
Name: Risk_Estrogen_Deficiency, dtype: int64

Risk_Immobilization
False    3424
Name: Risk_Immobilization, dtype: int64

Risk_Recurring_Falls
False    3424
Name: Risk_Recurring_Falls, dtype: int64

Count_Of_Risks
False    3424
Name: Count_Of_Risks, dtype: int64
```

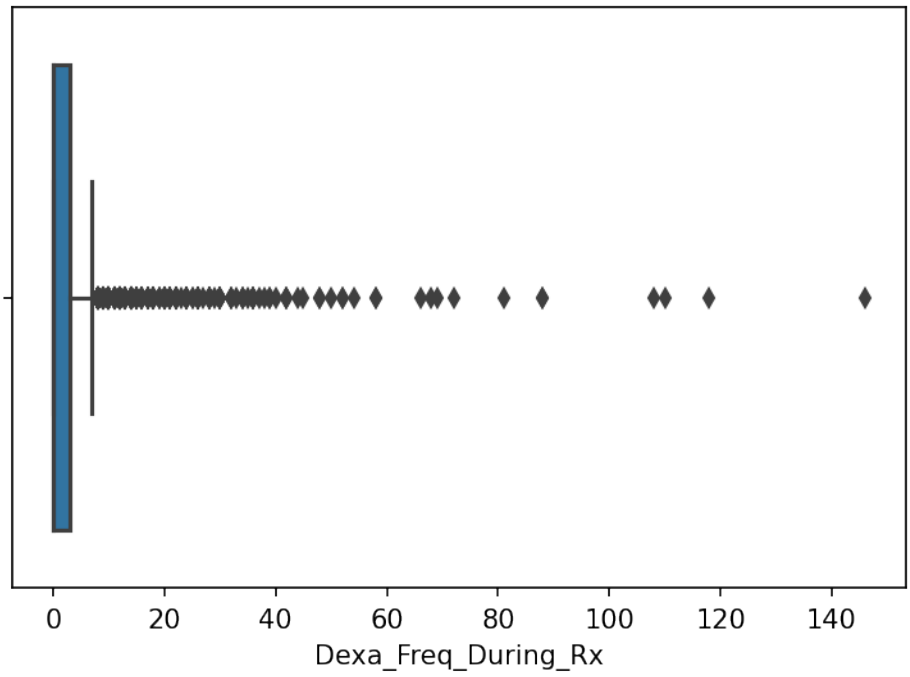
There is no missing data in the dataset

1.1.5 Checking for Outliers

Numerical values Lets visualize the column 'Dexa_Freq_During_Rx' with a box plot

[7]:
```
plt.figure(figsize=(6,4),dpi=150)
sns.boxplot(x=df['Dexa_Freq_During_Rx'])
```

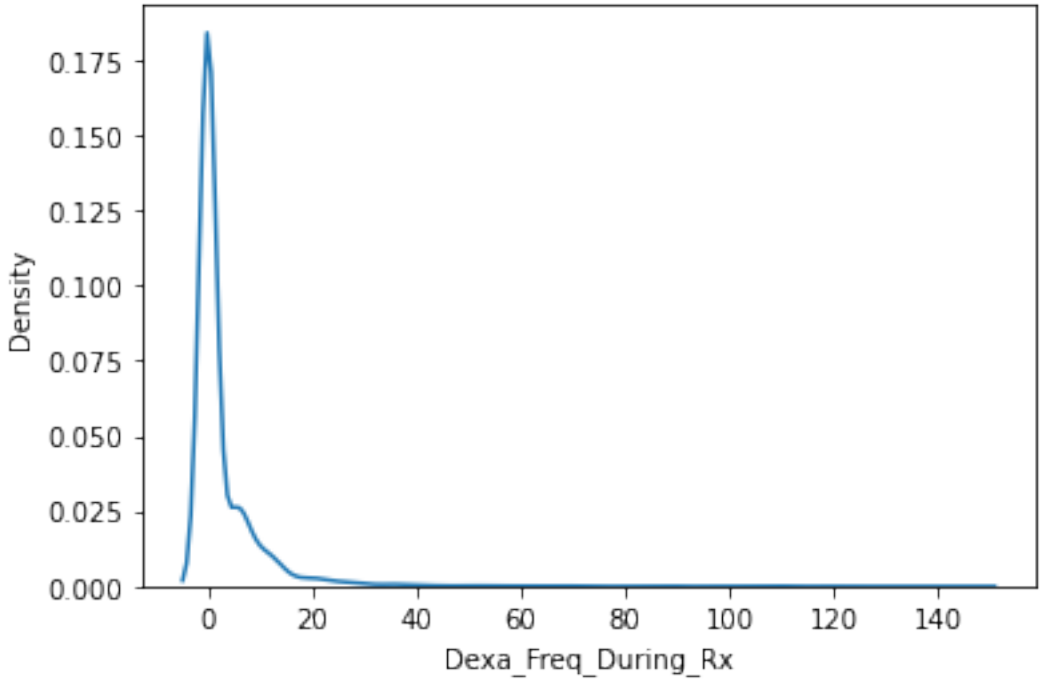
[7]:
```
<AxesSubplot:xlabel='Dexa_Freq_During_Rx'>
```



We can see there's a couple of outliers between 10 and 150

[8]:
```
sns.kdeplot(x=df["Dexa_Freq_During_Rx"])
```

[8]:
```
<AxesSubplot:xlabel='Dexa_Freq_During_Rx', ylabel='Density'>
```



The data is heavily positively skewed. We shall use the .skew() function to find out the exact extent.

[9]:
```
print(skew(df['Dexa_Freq_During_Rx']))
```

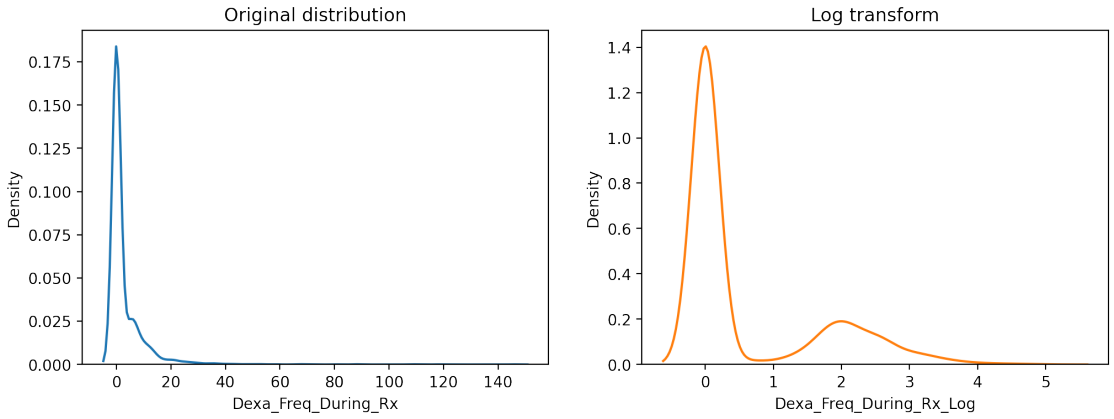
```
6.805747051718919
```

We shall apply log transformation to deal with this and replace the column with the log-transformmed version

[10]:
```
df["Dexa_Freq_During_Rx_Log"] = df['Dexa_Freq_During_Rx'].apply(lambda x: np.
↪log(1+x))
```

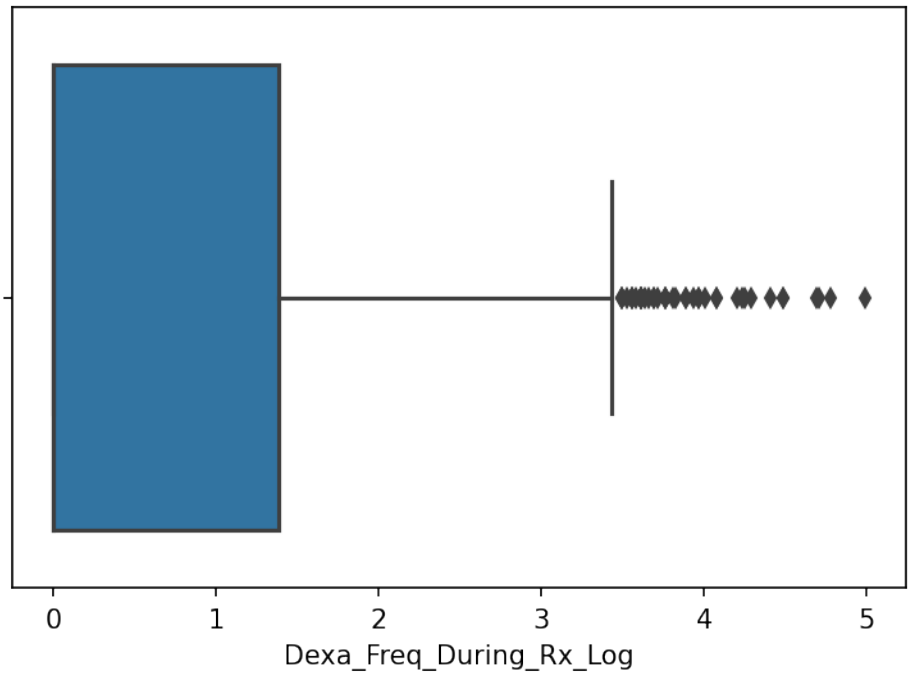
Let's compare these two columns' kde plots side by side

[11]:
```
fig,axes=plt.subplots(nrows=1,ncols=2)
fig.set_size_inches((12,4))
fig.set_dpi(200)
sns.kdeplot(x=df["Dexa_Freq_During_Rx"],ax=axes[0],color="tab:blue")
sns.kdeplot(x=df["Dexa_Freq_During_Rx_Log"],color="tab:orange");
axes[0].set_title("Original distribution")
axes[1].set_title("Log transform");
```



[12]:
```
#Box plot for the log transform data
plt.figure(figsize=(6,4),dpi=150)
sns.boxplot(x=df['Dexa_Freq_During_Rx_Log'])
```

[12]:
```
<AxesSubplot:xlabel='Dexa_Freq_During_Rx_Log'>
```



[13]:
```
print(skew(df['Dexa_Freq_During_Rx_Log']))
```

```
1.4052436284675567
```

The skewness has been greatly improved using log transformation

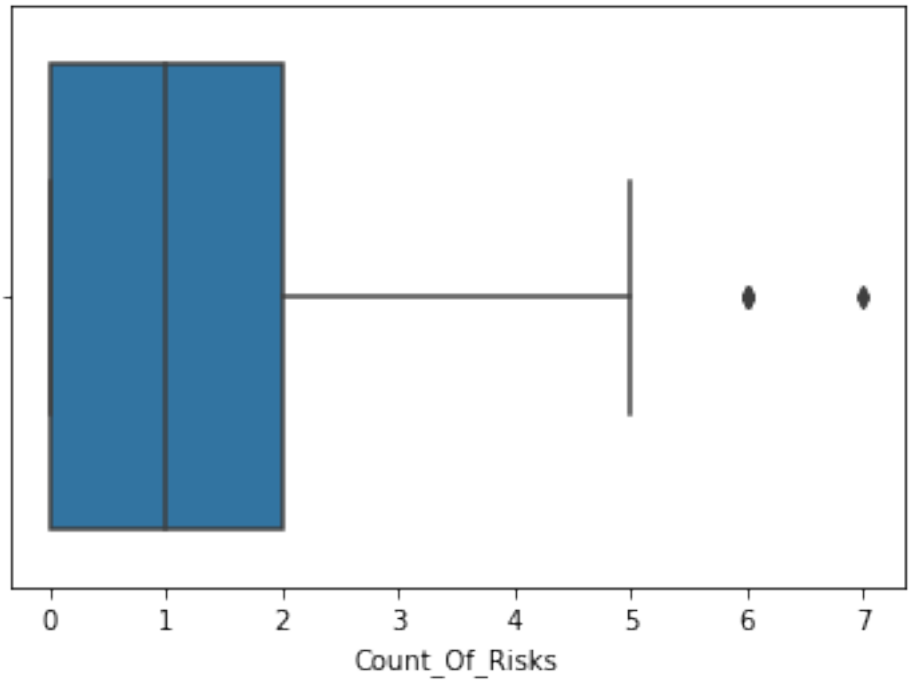
[14]:
```
print(skew(df['Count_Of_Risks']))
```

```
0.8794050541279611
```

A posite skewness is being observed. Lets visualize this column

[15]:
```
sns.boxplot(x=df['Count_Of_Risks'])
```

[15]:
```
<AxesSubplot:xlabel='Count_Of_Risks'>
```



Let's try replacing our outliers with the median

[16]:
```
median = df.loc[df['Count_Of_Risks'] <5, 'Count_Of_Risks'].median()
df.loc[df.Count_Of_Risks > 5, 'Count_Of_Risks'] = np.nan
df.fillna(median,inplace=True)
print(skew(df['Count_Of_Risks']))
```

```
0.7359181096502345
```

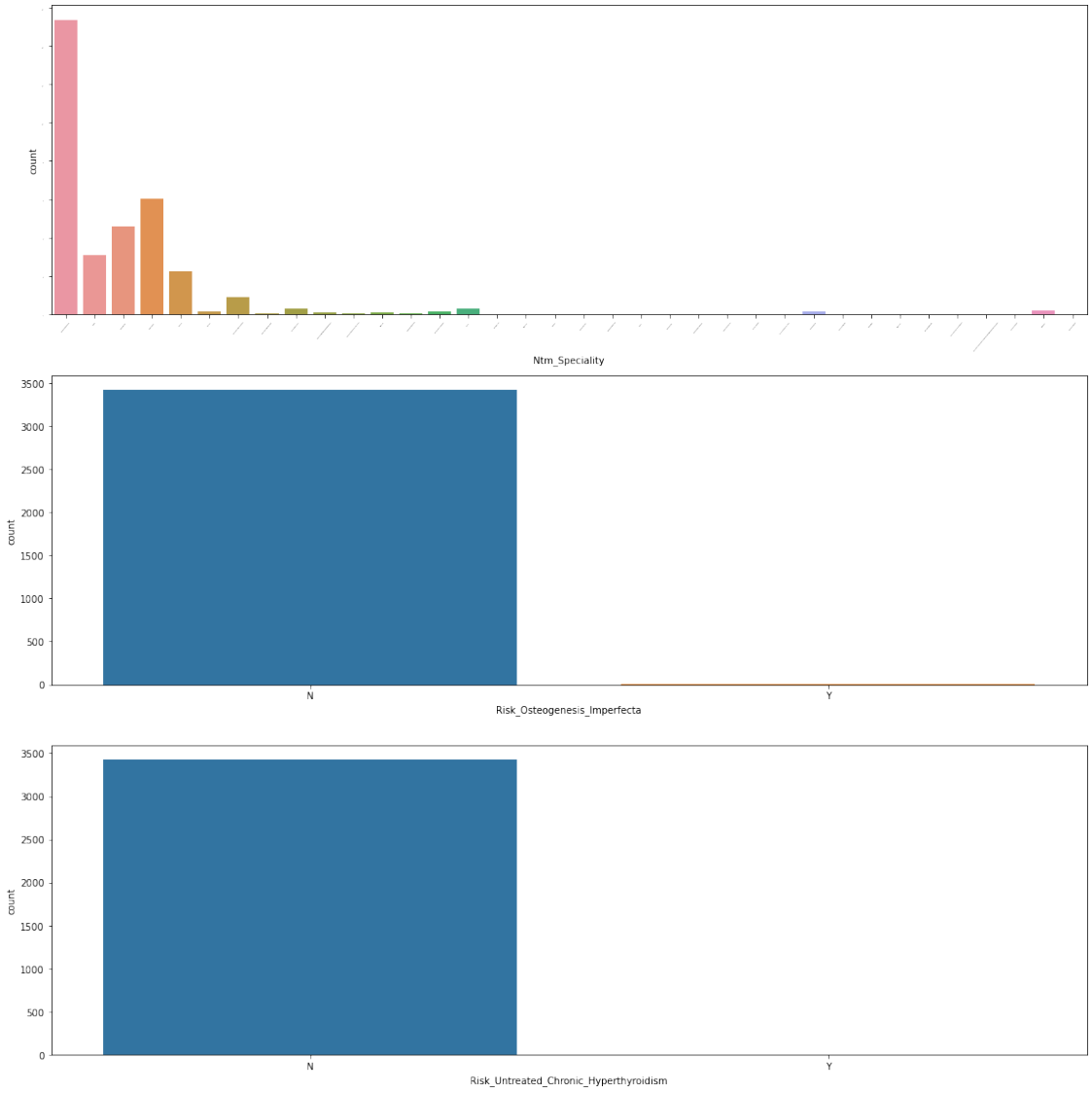
skewness has been reduce a little by replacing the outlier values with the median

Categorical values We will detect outliers by finding categories that have low frequencies with the help of histograms

[17]:
```
#Create a list of categorical columns
cat_cols=df.select_dtypes("object").drop("Ptid",axis=1).columns
#Create a list of categorical columns with outliers
cat_cols_outliers = cat_cols[[any(df[col].value_counts()<=10) for col in␣
 ↪cat_cols]]
```

[18]:
```
#Visualize the imbalance of categorical columns with outliers
fig,axes=plt.subplots(nrows=len(cat_cols_outliers))
fig.set_size_inches((16,4*4))
i=0
for col in cat_cols_outliers:
    sns.countplot(x=df[col],ax=axes[i])
```

```
    i+=1
axes[0].tick_params(rotation=50,labelsize=0)
plt.tight_layout()
```



Now let's retain these categogories and see how it'll affect the model

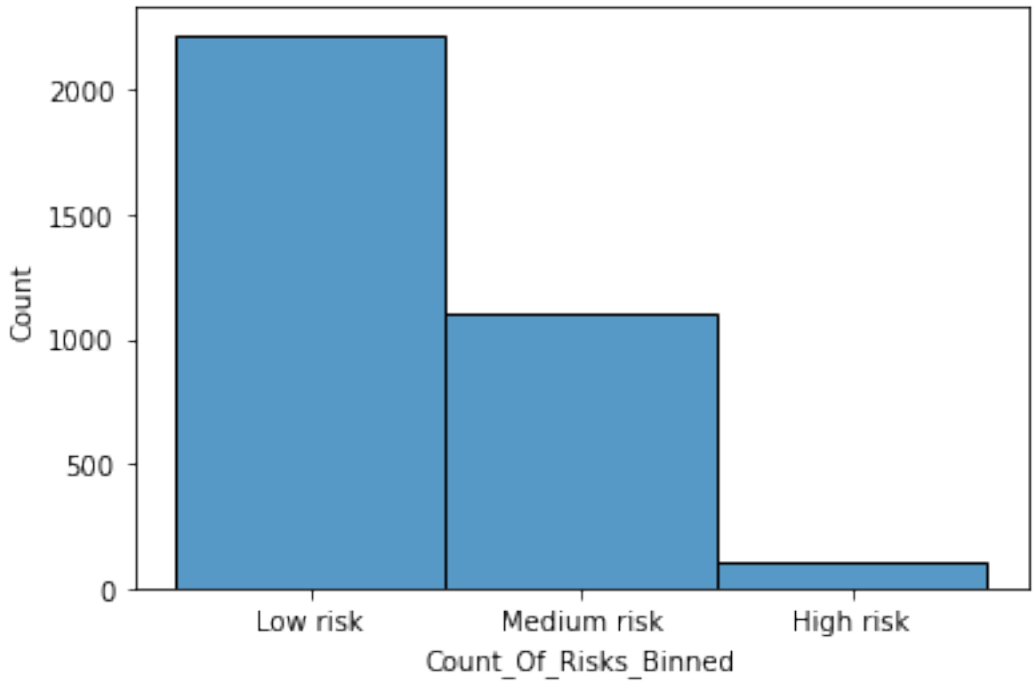
1.2 Feature scaling and transformation

1.2.1 Numerical values

We Have already scaled our 'Dexa_Freq_During_Rx_Log' using log Transformation in order to reduce outliers and minimize skewness. For the attribute "Count_Of_Risks", we will categorize our values in the following bins: Low Risk, Medium Risk and High Risk

[19]:
```
Risk_Bins = np.linspace(min(df['Count_Of_Risks']), max(df['Count_Of_Risks']), 4)
group_names = ["Low risk", "Medium risk", "High risk"]
df['Count_Of_Risks_Binned'] = pd.cut(df['Count_Of_Risks'], Risk_Bins, labels =␣
 ↪group_names, include_lowest = True)
sns.histplot(x=df['Count_Of_Risks_Binned'])
```

[19]:
```
<AxesSubplot:xlabel='Count_Of_Risks_Binned', ylabel='Count'>
```



1.2.2 Categorical values

We have both nominal and ordinal data in our dataset. We will be using ordinal encoding on our ordinal data and frequency encoding on our noiminal data.

[20]:
```
print(cat_cols)
```

```
Index(['Persistency_Flag', 'Gender', 'Race', 'Ethnicity', 'Region',
       'Age_Bucket', 'Ntm_Speciality', 'Ntm_Specialist_Flag',
       'Ntm_Speciality_Bucket', 'Gluco_Record_Prior_Ntm',
       'Gluco_Record_During_Rx', 'Dexa_During_Rx', 'Frag_Frac_Prior_Ntm',
       'Frag_Frac_During_Rx', 'Risk_Segment_Prior_Ntm',
       'Tscore_Bucket_Prior_Ntm', 'Risk_Segment_During_Rx',
       'Tscore_Bucket_During_Rx', 'Change_T_Score', 'Change_Risk_Segment',
       'Adherent_Flag', 'Idn_Indicator', 'Injectable_Experience_During_Rx',
       'Comorb_Encounter_For_Screening_For_Malignant_Neoplasms',
       'Comorb_Encounter_For_Immunization',
```

```
       'Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx',
       'Comorb_Vitamin_D_Deficiency',
       'Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified',
       'Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx',
       'Comorb_Long_Term_Current_Drug_Therapy', 'Comorb_Dorsalgia',
       'Comorb_Personal_History_Of_Other_Diseases_And_Conditions',
       'Comorb_Other_Disorders_Of_Bone_Density_And_Structure',
       'Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias',
       'Comorb_Osteoporosis_without_current_pathological_fracture',
       'Comorb_Personal_history_of_malignant_neoplasm',
       'Comorb_Gastro_esophageal_reflux_disease',
       'Concom_Cholesterol_And_Triglyceride_Regulating_Preparations',
       'Concom_Narcotics', 'Concom_Systemic_Corticosteroids_Plain',
       'Concom_Anti_Depressants_And_Mood_Stabilisers',
       'Concom_Fluoroquinolones', 'Concom_Cephalosporins',
       'Concom_Macrolides_And_Similar_Types',
       'Concom_Broad_Spectrum_Penicillins', 'Concom_Anaesthetics_General',
       'Concom_Viral_Vaccines', 'Risk_Type_1_Insulin_Dependent_Diabetes',
       'Risk_Osteogenesis_Imperfecta', 'Risk_Rheumatoid_Arthritis',
       'Risk_Untreated_Chronic_Hyperthyroidism',
       'Risk_Untreated_Chronic_Hypogonadism', 'Risk_Untreated_Early_Menopause',
       'Risk_Patient_Parent_Fractured_Their_Hip', 'Risk_Smoking_Tobacco',
       'Risk_Chronic_Malnutrition_Or_Malabsorption',
       'Risk_Chronic_Liver_Disease', 'Risk_Family_History_Of_Osteoporosis',
       'Risk_Low_Calcium_Intake', 'Risk_Vitamin_D_Insufficiency',
       'Risk_Poor_Health_Frailty', 'Risk_Excessive_Thinness',
       'Risk_Hysterectomy_Oophorectomy', 'Risk_Estrogen_Deficiency',
       'Risk_Immobilization', 'Risk_Recurring_Falls'],
      dtype='object')
```

After going through the dataset, I have found the following ordinal columns: - 'Age_Bucket' - 'Tscore_Bucket_Prior_Ntm' - 'Tscore_Bucket_During_Rx'

[21]:
```
# Importing ordinal encoder
data_categorical = df[cat_cols]
from sklearn.preprocessing import OrdinalEncoder
Age_column = data_categorical[["Age_Bucket"]]
encoder = OrdinalEncoder()
Age_encoded = encoder.fit_transform(Age_column)
Age_encoded
```

[21]:
```
array([[3.],
       [0.],
       [1.],
       …,
       [3.],
       [0.],
       [1.]])
```

[22]:
```
Tscore_Bucket_Prior_Ntm_column = data_categorical[["Tscore_Bucket_Prior_Ntm"]]
encoder = OrdinalEncoder()
Tscore_Bucket_Prior_Ntm_encoded = encoder.
 ↪fit_transform(Tscore_Bucket_Prior_Ntm_column)
Tscore_Bucket_Prior_Ntm_encoded
```

[22]:
```
array([[1.],
       [1.],
       [0.],
       …,
       [1.],
       [1.],
       [1.]])
```

[23]:
```
Tscore_Bucket_During_Rx_column = data_categorical[["Tscore_Bucket_During_Rx"]]
encoder = OrdinalEncoder()
Tscore_Bucket_During_Rx_encoded = encoder.
 ↪fit_transform(Tscore_Bucket_During_Rx_column)
Tscore_Bucket_During_Rx_encoded
```

[23]:
```
array([[0.],
       [2.],
       [0.],
       …,
       [0.],
       [2.],
       [2.]])
```

[24]:
```
df["Age_encoded"] = Age_encoded
df["Tscore_Bucket_Prior_Ntm_encoded"] = Tscore_Bucket_Prior_Ntm_encoded
df["Tscore_Bucket_During_Rx_encoded"] = Tscore_Bucket_During_Rx_encoded
```

Now for the Nominal colums

[25]:
```
df_nominal = data_categorical.drop(["Tscore_Bucket_During_Rx",␣
 ↪"Tscore_Bucket_Prior_Ntm", "Age_Bucket"], axis = 1)
```

[26]:
```
# Using frequency encoding
for column in df_nominal:
    Freq_enc = (df_nominal.groupby(column).size()) / len(df)
    print(Freq_enc)
```

```
Persistency_Flag
Non-Persistent    0.62354
Persistent        0.37646
dtype: float64
Gender
Female    0.943341
```

```
Male      0.056659
dtype: float64
Race
African American    0.027745
Asian               0.024533
Caucasian           0.919393
Other/Unknown       0.028329
dtype: float64
Ethnicity
Hispanic        0.028621
Not Hispanic    0.944801
Unknown         0.026577
dtype: float64
Region
Midwest          0.403914
Northeast        0.067757
Other/Unknown    0.017523
South            0.364194
West             0.146612
dtype: float64
Ntm_Speciality
CARDIOLOGY                                                          0.006425
CLINICAL NURSE SPECIALIST                                           0.000292
EMERGENCY MEDICINE                                                  0.000292
ENDOCRINOLOGY                                                       0.133762
GASTROENTEROLOGY                                                    0.000584
GENERAL PRACTITIONER                                                0.448306
GERIATRIC MEDICINE                                                  0.000584
HEMATOLOGY & ONCOLOGY                                               0.004089
HOSPICE AND PALLIATIVE MEDICINE                                     0.000584
HOSPITAL MEDICINE                                                   0.000292
NEPHROLOGY                                                          0.000876
NEUROLOGY                                                           0.000292
NUCLEAR MEDICINE                                                    0.000292
OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY      0.000292
OBSTETRICS AND GYNECOLOGY                                           0.026285
OCCUPATIONAL MEDICINE                                               0.000292
ONCOLOGY                                                            0.065713
OPHTHALMOLOGY                                                       0.000292
ORTHOPEDIC SURGERY                                                  0.008762
ORTHOPEDICS                                                         0.000876
OTOLARYNGOLOGY                                                      0.004089
PAIN MEDICINE                                                       0.000292
PATHOLOGY                                                           0.004673
PEDIATRICS                                                          0.003797
PHYSICAL MEDICINE AND REHABILITATION                                0.003213
PLASTIC SURGERY                                                     0.000584
PODIATRY                                                            0.000292
```

```
PSYCHIATRY AND NEUROLOGY                                    0.001168
PULMONARY MEDICINE                                          0.002336
RADIOLOGY                                                   0.000292
RHEUMATOLOGY                                                0.176402
SURGERY AND SURGICAL SPECIALTIES                            0.002336
TRANSPLANT SURGERY                                          0.000584
UROLOGY                                                     0.009638
Unknown                                                     0.090537
VASCULAR SURGERY                                            0.000584
dtype: float64
Ntm_Specialist_Flag
Others        0.587909
Specialist    0.412091
dtype: float64
Ntm_Speciality_Bucket
Endo/Onc/Uro                  0.209112
OB/GYN/Others/PCP/Unknown     0.614486
Rheum                         0.176402
dtype: float64
Gluco_Record_Prior_Ntm
N    0.764895
Y    0.235105
dtype: float64
Gluco_Record_During_Rx
N    0.736565
Y    0.263435
dtype: float64
Dexa_During_Rx
N    0.726636
Y    0.273364
dtype: float64
Frag_Frac_Prior_Ntm
N    0.838785
Y    0.161215
dtype: float64
Frag_Frac_During_Rx
N    0.878213
Y    0.121787
dtype: float64
Risk_Segment_Prior_Ntm
HR_VHR    0.43604
VLR_LR    0.56396
dtype: float64
Risk_Segment_During_Rx
HR_VHR     0.281834
Unknown    0.437208
VLR_LR     0.280958
dtype: float64
```

```
Change_T_Score
Improved     0.027453
No change    0.484813
Unknown      0.437208
Worsened     0.050526
dtype: float64
Change_Risk_Segment
Improved     0.006425
No change    0.307243
Unknown      0.650993
Worsened     0.035339
dtype: float64
Adherent_Flag
Adherent        0.949474
Non-Adherent    0.050526
dtype: float64
Idn_Indicator
N    0.253213
Y    0.746787
dtype: float64
Injectable_Experience_During_Rx
N    0.107477
Y    0.892523
dtype: float64
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms
N    0.552278
Y    0.447722
dtype: float64
Comorb_Encounter_For_Immunization
N    0.558119
Y    0.441881
dtype: float64
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx
N    0.60514
Y    0.39486
dtype: float64
Comorb_Vitamin_D_Deficiency
N    0.680783
Y    0.319217
dtype: float64
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified
N    0.708236
Y    0.291764
dtype: float64
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx
N    0.768984
Y    0.231016
dtype: float64
```

```
Comorb_Long_Term_Current_Drug_Therapy
N    0.76139
Y    0.23861
dtype: float64
Comorb_Dorsalgia
N    0.772488
Y    0.227512
dtype: float64
Comorb_Personal_History_Of_Other_Diseases_And_Conditions
N    0.802278
Y    0.197722
dtype: float64
Comorb_Other_Disorders_Of_Bone_Density_And_Structure
N    0.848715
Y    0.151285
dtype: float64
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias
N    0.484521
Y    0.515479
dtype: float64
Comorb_Osteoporosis_without_current_pathological_fracture
N    0.732185
Y    0.267815
dtype: float64
Comorb_Personal_history_of_malignant_neoplasm
N    0.810456
Y    0.189544
dtype: float64
Comorb_Gastro_esophageal_reflux_disease
N    0.816005
Y    0.183995
dtype: float64
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations
N    0.65479
Y    0.34521
dtype: float64
Concom_Narcotics
N    0.639895
Y    0.360105
dtype: float64
Concom_Systemic_Corticosteroids_Plain
N    0.715829
Y    0.284171
dtype: float64
Concom_Anti_Depressants_And_Mood_Stabilisers
N    0.719918
Y    0.280082
dtype: float64
```

```
Concom_Fluoroquinolones
N    0.81396
Y    0.18604
dtype: float64
Concom_Cephalosporins
N    0.82389
Y    0.17611
dtype: float64
Concom_Macrolides_And_Similar_Types
N    0.833236
Y    0.166764
dtype: float64
Concom_Broad_Spectrum_Penicillins
N    0.871787
Y    0.128213
dtype: float64
Concom_Anaesthetics_General
N    0.854848
Y    0.145152
dtype: float64
Concom_Viral_Vaccines
N    0.896904
Y    0.103096
dtype: float64
Risk_Type_1_Insulin_Dependent_Diabetes
N    0.959404
Y    0.040596
dtype: float64
Risk_Osteogenesis_Imperfecta
N    0.999124
Y    0.000876
dtype: float64
Risk_Rheumatoid_Arthritis
N    0.962033
Y    0.037967
dtype: float64
Risk_Untreated_Chronic_Hyperthyroidism
N    0.999416
Y    0.000584
dtype: float64
Risk_Untreated_Chronic_Hypogonadism
N    0.962909
Y    0.037091
dtype: float64
Risk_Untreated_Early_Menopause
N    0.996495
Y    0.003505
dtype: float64
```

```
Risk_Patient_Parent_Fractured_Their_Hip
N    0.925234
Y    0.074766
dtype: float64
Risk_Smoking_Tobacco
N    0.811916
Y    0.188084
dtype: float64
Risk_Chronic_Malnutrition_Or_Malabsorption
N    0.862734
Y    0.137266
dtype: float64
Risk_Chronic_Liver_Disease
N    0.994743
Y    0.005257
dtype: float64
Risk_Family_History_Of_Osteoporosis
N    0.895444
Y    0.104556
dtype: float64
Risk_Low_Calcium_Intake
N    0.987734
Y    0.012266
dtype: float64
Risk_Vitamin_D_Insufficiency
N    0.522196
Y    0.477804
dtype: float64
Risk_Poor_Health_Frailty
N    0.943925
Y    0.056075
dtype: float64
Risk_Excessive_Thinness
N    0.980432
Y    0.019568
dtype: float64
Risk_Hysterectomy_Oophorectomy
N    0.984229
Y    0.015771
dtype: float64
Risk_Estrogen_Deficiency
N    0.996787
Y    0.003213
dtype: float64
Risk_Immobilization
N    0.995911
Y    0.004089
dtype: float64
```

```
Risk_Recurring_Falls
N    0.979848
Y    0.020152
dtype: float64
```

[27]:
```
df.head(5)
```

[27]:

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region
0	P1	Persistent	Male	Caucasian	Not Hispanic	West
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest

	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag
0	>75	GENERAL PRACTITIONER	Others
1	55-65	GENERAL PRACTITIONER	Others
2	65-75	GENERAL PRACTITIONER	Others
3	>75	GENERAL PRACTITIONER	Others
4	>75	GENERAL PRACTITIONER	Others

	Ntm_Speciality_Bucket	…	Risk_Hysterectomy_Oophorectomy
0	OB/GYN/Others/PCP/Unknown	…	N
1	OB/GYN/Others/PCP/Unknown	…	N
2	OB/GYN/Others/PCP/Unknown	…	N
3	OB/GYN/Others/PCP/Unknown	…	N
4	OB/GYN/Others/PCP/Unknown	…	N

	Risk_Estrogen_Deficiency	Risk_Immobilization	Risk_Recurring_Falls
0	N	N	N
1	N	N	N
2	N	N	N
3	N	N	N
4	N	N	N

	Count_Of_Risks	Dexa_Freq_During_Rx_Log	Count_Of_Risks_Binned	Age_encoded
0	0.0	0.0	Low risk	3.0
1	0.0	0.0	Low risk	0.0
2	2.0	0.0	Medium risk	1.0
3	1.0	0.0	Low risk	3.0
4	1.0	0.0	Low risk	3.0

	Tscore_Bucket_Prior_Ntm_encoded	Tscore_Bucket_During_Rx_encoded
0	1.0	0.0
1	1.0	2.0
2	0.0	0.0
3	1.0	0.0

```
4                              0.0                              2.0

[5 rows x 74 columns]
```