Data Science Intern at Data Glacier

Project: Healthcare - Persistency of a drug

Week 9: Deliverables

Name: Amraoui Fatima Ezzahra

Email: amraouifatimaezzahra@gmail.com

Country: Morocco

Specialization: Data Science

Batch Code: LISUM13

Date: 2 November 2022

Submitted to: Data Glacier

Table of Contents:

1. Problem Description

Persistence represents the time (e.g., days, months, years) over which a patient continues the treatment. For practical reasons, it might be assessed according to the time taken for a patient to fill their prescription and can capture both the timeliness and frequency of refilling. In reality, as defined by the adherence taxonomy, adherence is a dynamic behavior, consisting of initiation, implementation and discontinuation phases of treatment that vary over time, resulting in periods of persistence and non-persistence. Therefore, rather than measuring the specific components of adherence, we could measure persistence, which captures the chronology of adherence and enables us to examine and understand patterns of medication-taking behavior.
For that this project aims to build an automated classifier that predict whether a patient was persistent or not.

2. Data understanding

For dataset I choose to work on healthcare – persistency of a drug dataset, that it's available in canvas. The dataset is contains 3424 rows × 69 columns. The attributes are presented below:

Variable	Variable Description
Patient ID	Unique ID of each patient
Persistency_Flag	Flag indicating if a patient was persistent or not
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
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)
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 (>=7.5mg 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
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

The dataframe is presented as follow:

```
df = pd.read_excel(xls, 'Dataset', index_col= None)
df
```

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...	Risk_Family_History_Of_Osteoporosis	Risk_Low_Calci
0	P1	Persistent	Male	Caucasian	Not Hispanic	West	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	N	
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West	55-65	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	N	
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest	65-75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	N	
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	N	
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	N	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3419	P3420	Persistent	Female	Caucasian	Not Hispanic	South	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	N	
3420	P3421	Persistent	Female	Caucasian	Not Hispanic	South	>75	Unknown	Others	OB/GYN/Others/PCP/Unknown	...	N	
3421	P3422	Persistent	Female	Caucasian	Not Hispanic	South	>75	ENDOCRINOLOGY	Specialist	Endo/Onc/Uro	...	N	
3422	P3423	Non-Persistent	Female	Caucasian	Not Hispanic	South	55-65	Unknown	Others	OB/GYN/Others/PCP/Unknown	...	N	
3423	P3424	Non-Persistent	Female	Caucasian	Not Hispanic	South	65-75	Unknown	Others	OB/GYN/Others/PCP/Unknown	...	N	

3424 rows × 69 columns

The dataset contains 69 variables :

```
[ ]  df.columns
```

```
Index(['Ptid', '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_Freq_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', 'Count_Of_Risks'],
      dtype='object')
```

With the Dependent variable is : Persistency_flag

The dataset has only 2 quantitative variables and the rest is qualitative(nominal):

```
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
```

3. Data preprocessing

First let's check if we have some missing values:

```
[ ] df.isna()
```

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...	Risk_Family_History_Of_Osteoporosis	Risk_Low_Calcium_Intake	Risk_Vitamin_D_Insufficiency	Ri
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3419	False	False	False	False	False	False	False	False	False	False	...	False	False	False	
3420	False	False	False	False	False	False	False	False	False	False	...	False	False	False	
3421	False	False	False	False	False	False	False	False	False	False	...	False	False	False	
3422	False	False	False	False	False	False	False	False	False	False	...	False	False	False	
3423	False	False	False	False	False	False	False	False	False	False	...	False	False	False	

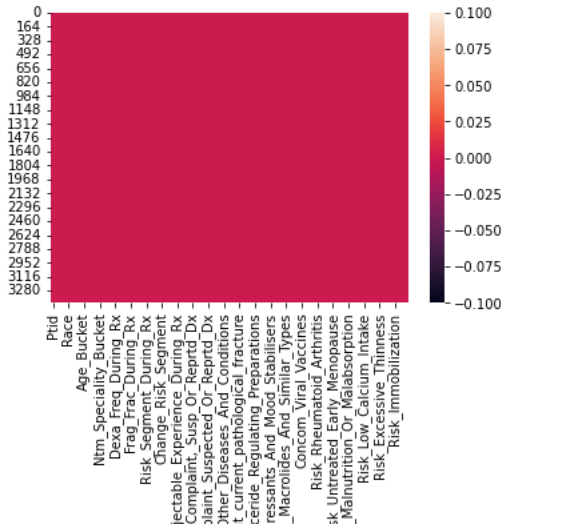
3424 rows × 69 columns

```
[ ] df.isna().sum()
```

```
Ptid                                0
Persistency_Flag                    0
Gender                              0
Race                                0
Ethnicity                           0
                                   ..
Risk_Hysterectomy_Oophorectomy      0
Risk_Estrogen_Deficiency            0
Risk_Immobilization                 0
Risk_Recurring_Falls                0
Count_Of_Risks                      0
Length: 69, dtype: int64
```

```
[ ] sns.heatmap(df.isna())
```

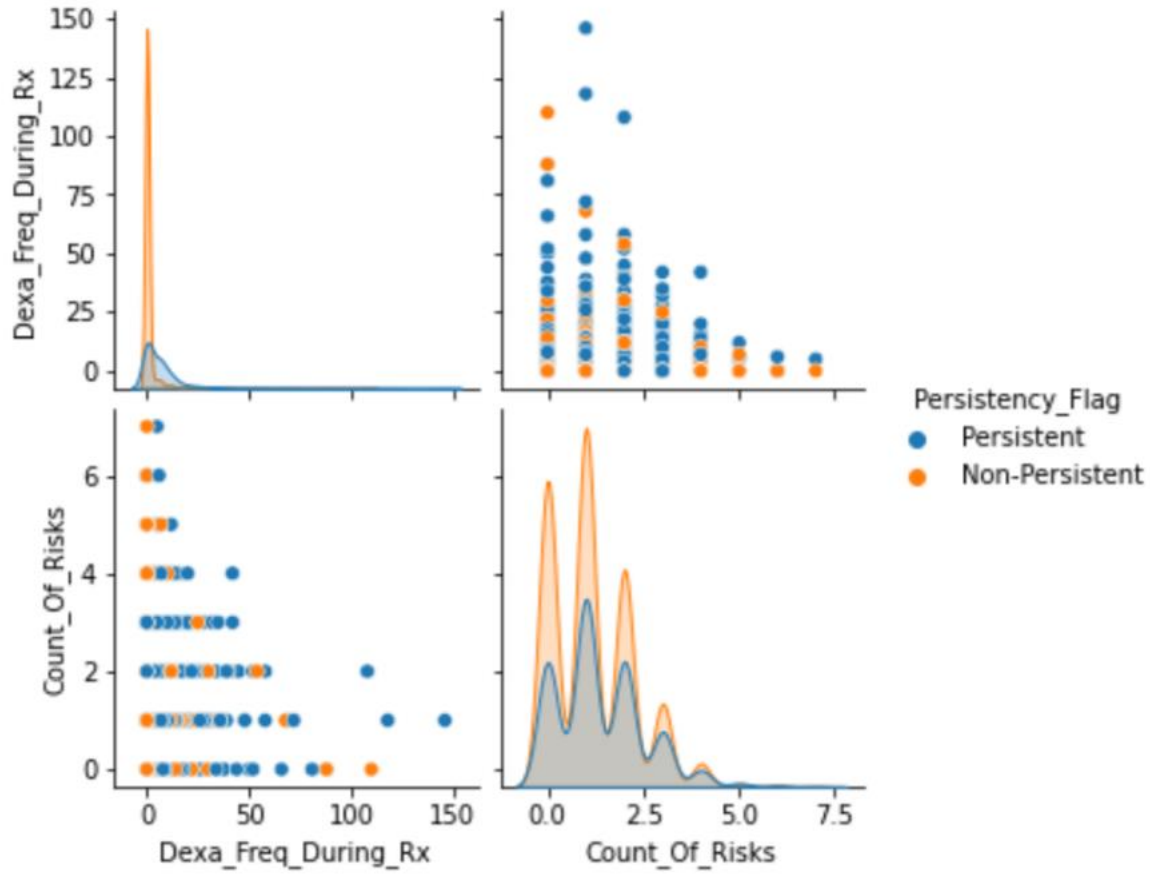
```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff4480dbc90>
```



We see that we don’t have missing values which is good for our analyses, for outliers since all the attributes are nominal except 2 variables we are not interested about outliers when having categorical data .
For numerical var :

```
[ ]  sns.pairplot(data=df, hue='Persistency_Flag')
```

```
<seaborn.axisgrid.PairGrid at 0x7ff4438ae450>
```



We notice that Dexa_freq_dunning_Rx is relatively higher for people who's not following a treatment than others who's pursuing a drug .
Same remarque for count_of_risks which is greater for people that are not persistent compared to Persistent patient.

Now let's get more about our categorical data by displaying different categories or groups for each Variable :

```
-Column name is: Persistency_Flag and it value is: Non-Persistent    2135
Persistent        1289
Name: Persistency_Flag, dtype: int64
-Column name is: Gender and it value is: Female    3230
Male       194
Name: Gender, dtype: int64
-Column name is: Race and it value is: Caucasian           3148
Other/Unknown         97
African American      95
Asian                 84
Name: Race, dtype: int64
-Column name is: Ethnicity and it value is: Not Hispanic    3235
Hispanic          98
Unknown           91
Name: Ethnicity, dtype: int64
-Column name is: Region and it value is: Midwest          1383
South            1247
West              502
Northeast         232
Other/Unknown      60
Name: Region, dtype: int64
-Column name is: Age_Bucket and it value is: >75      1439
65-75    1086
55-65     733
<55       166
Name: Age_Bucket, dtype: int64
-Column name is: Ntm_Speciality and it value is: GENERAL PRACTITIONER
RHEUMATOLOGY                                                    604
ENDOCRINOLOGY                                                   458
Unknown                                                         310
ONCOLOGY                                                        225
OBSTETRICS AND GYNECOLOGY                                        90
UROLOGY                                                          33
ORTHOPEDIC SURGERY                                               30
CARDIOLOGY                                                       22
PATHOLOGY                                                        16
HEMATOLOGY & ONCOLOGY                                            14
OTOLARYNGOLOGY                                                   14
PEDIATRICS                                                       13
PHYSICAL MEDICINE AND REHABILITATION                             11
PULMONARY MEDICINE                                                8
SURGERY AND SURGICAL SPECIALTIES                                  8
SURGERY AND SURGICAL SPECIALTIES                                  8
PSYCHIATRY AND NEUROLOGY                                          4
NEPHROLOGY                                                        3
ORTHOPEDICS                                                       3
PLASTIC SURGERY                                                   2
VASCULAR SURGERY                                                  2
HOSPICE AND PALLIATIVE MEDICINE                                   2
GERIATRIC MEDICINE                                                2
GASTROENTEROLOGY                                                  2
TRANSPLANT SURGERY                                                2
CLINICAL NURSE SPECIALIST                                         1
OCCUPATIONAL MEDICINE                                             1
HOSPITAL MEDICINE                                                 1
OPHTHALMOLOGY                                                     1
PODIATRY                                                          1
EMERGENCY MEDICINE                                                1
RADIOLOGY                                                         1
OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY    1
NEUROLOGY                                                         1
PAIN MEDICINE                                                     1
NUCLEAR MEDICINE                                                  1
Name: Ntm_Speciality, dtype: int64
-Column name is: Ntm_Specialist_Flag and it value is: Others        2013
Specialist    1411
Name: Ntm_Specialist_Flag, dtype: int64
-Column name is: Ntm_Speciality_Bucket and it value is: OB/GYN/Others/PCP/Unknown    2104
Endo/Onc/Uro                716
Rheum                       604
Name: Ntm_Speciality_Bucket, dtype: int64
-Column name is: Gluco_Record_Prior_Ntm and it value is: N    2619
Y     805
Name: Gluco_Record_Prior_Ntm, dtype: int64
-Column name is: Gluco_Record_During_Rx and it value is: N    2522
Y     902
Name: Gluco_Record_During_Rx, dtype: int64
-Column name is: Dexa_During_Rx and it value is: N    2488
Y     936
Name: Dexa_During_Rx, dtype: int64
-Column name is: Frag_Frac_Prior_Ntm and it value is: N    2872
Y     552
Name: Frag_Frac_Prior_Ntm, dtype: int64
-Column name is: Frag_Frac_During_Rx and it value is: N    3007
```

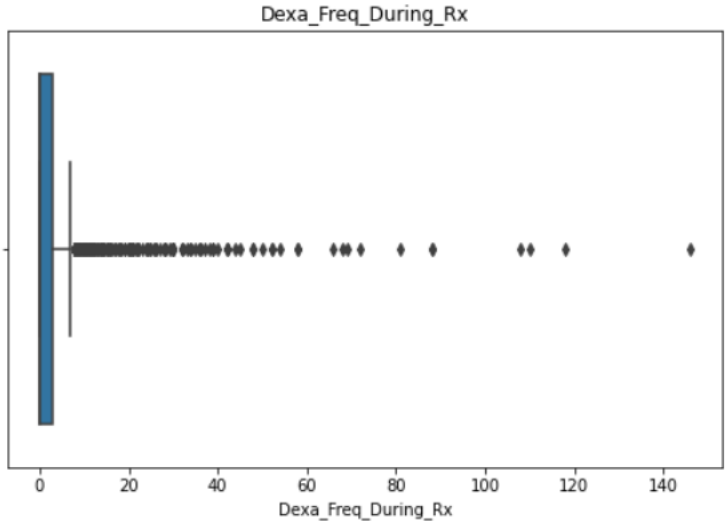
```
Y     417
Name: Frag_Frac_During_Rx, dtype: int64
-Column name is: Risk_Segment_Prior_Ntm and it value is: VLR_LR    1931
HR_VHR    1493
Name: Risk_Segment_Prior_Ntm, dtype: int64
-Column name is: Tscore_Bucket_Prior_Ntm and it value is: >-2.5     1951
<=-2.5    1473
Name: Tscore_Bucket_Prior_Ntm, dtype: int64
-Column name is: Risk_Segment_During_Rx and it value is: Unknown    1497
HR_VHR      965
VLR_LR      962
Name: Risk_Segment_During_Rx, dtype: int64
-Column name is: Tscore_Bucket_During_Rx and it value is: Unknown    1497
<=-2.5     1017
>-2.5       910
Name: Tscore_Bucket_During_Rx, dtype: int64
-Column name is: Change_T_Score and it value is: No change    1660
Unknown      1497
Worsened      173
Improved       94
Name: Change_T_Score, dtype: int64
-Column name is: Change_Risk_Segment and it value is: Unknown      2229
No change    1052
Worsened      121
Improved       22
Name: Change_Risk_Segment, dtype: int64
-Column name is: Adherent_Flag and it value is: Adherent        3251
Non-Adherent     173
Name: Adherent_Flag, dtype: int64
-Column name is: Idn_Indicator and it value is: Y    2557
N     867
Name: Idn_Indicator, dtype: int64
-Column name is: Injectable_Experience_During_Rx and it value is: Y    3056
N     368
Name: Injectable_Experience_During_Rx, dtype: int64
-Column name is: Comorb_Encounter_For_Screening_For_Malignant_Neoplasms and it value is: N    1891
Y    1533
Name: Comorb_Encounter_For_Screening_For_Malignant_Neoplasms, dtype: int64
-Column name is: Comorb_Encounter_For_Immunization and it value is: N    1911
Y    1513
Name: Comorb_Encounter_For_Immunization, dtype: int64
-Column name is: Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx and it value is: N    2072
Y    1352
Name: Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx, dtype: int64
-Column name is: Comorb_Vitamin_D_Deficiency and it value is: N    2331
Y    1093
Name: Comorb_Vitamin_D_Deficiency, dtype: int64
-Column name is: Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified and it value is: N    2425
Y     999
Name: Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified, dtype: int64
-Column name is: Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx and it value is: N    2633
Y     791
Name: Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx, dtype: int64
-Column name is: Comorb_Long_Term_Current_Drug_Therapy and it value is: N    2607
Y     817
Name: Comorb_Long_Term_Current_Drug_Therapy, dtype: int64
-Column name is: Comorb_Dorsalgia and it value is: N    2645
Y     779
Name: Comorb_Dorsalgia, dtype: int64
-Column name is: Comorb_Personal_History_Of_Other_Diseases_And_Conditions and it value is: N    2747
Y     677
Name: Comorb_Personal_History_Of_Other_Diseases_And_Conditions, dtype: int64
-Column name is: Comorb_Other_Disorders_Of_Bone_Density_And_Structure and it value is: N    2906
Y     518
Name: Comorb_Other_Disorders_Of_Bone_Density_And_Structure, dtype: int64
-Column name is: Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias and it value is: Y    1765
N    1659
Name: Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias, dtype: int64
-Column name is: Comorb_Osteoporosis_without_current_pathological_fracture and it value is: N    2507
Y     917
Name: Comorb_Osteoporosis_without_current_pathological_fracture, dtype: int64
-Column name is: Comorb_Personal_history_of_malignant_neoplasm and it value is: N    2775
Y     649
Name: Comorb_Personal_history_of_malignant_neoplasm, dtype: int64
-Column name is: Comorb_Gastro_esophageal_reflux_disease and it value is: N    2794
Y     630
Name: Comorb_Gastro_esophageal_reflux_disease, dtype: int64
-Column name is: Concom_Cholesterol_And_Triglyceride_Regulating_Preparations and it value is: N    2242
Y    1182
Name: Concom_Cholesterol_And_Triglyceride_Regulating_Preparations, dtype: int64
-Column name is: Concom_Narcotics and it value is: N    2191
Y    1233
Name: Concom_Narcotics, dtype: int64
-Column name is: Concom_Systemic_Corticosteroids_Plain and it value is: N    2451
```

```
Y     973
Name: Concom_Systemic_Corticosteroids_Plain, dtype: int64
-Column name is: Concom_Anti_Depressants_And_Mood_Stabilisers and it value is: N    2465
Y     959
Name: Concom_Anti_Depressants_And_Mood_Stabilisers, dtype: int64
-Column name is: Concom_Fluoroquinolones and it value is: N    2787
Y     637
Name: Concom_Fluoroquinolones, dtype: int64
-Column name is: Concom_Cephalosporins and it value is: N    2821
Y     603
Name: Concom_Cephalosporins, dtype: int64
-Column name is: Concom_Macrolides_And_Similar_Types and it value is: N    2853
Y     571
Name: Concom_Macrolides_And_Similar_Types, dtype: int64
-Column name is: Concom_Broad_Spectrum_Penicillins and it value is: N    2985
Y     439
Name: Concom_Broad_Spectrum_Penicillins, dtype: int64
-Column name is: Concom_Anaesthetics_General and it value is: N    2927
Y     497
Name: Concom_Anaesthetics_General, dtype: int64
-Column name is: Concom_Viral_Vaccines and it value is: N    3071
Y     353
Name: Concom_Viral_Vaccines, dtype: int64
-Column name is: Risk_Type_1_Insulin_Dependent_Diabetes and it value is: N    3285
Y     139
Name: Risk_Type_1_Insulin_Dependent_Diabetes, dtype: int64
-Column name is: Risk_Osteogenesis_Imperfecta and it value is: N    3421
Y       3
Name: Risk_Osteogenesis_Imperfecta, dtype: int64
-Column name is: Risk_Rheumatoid_Arthritis and it value is: N    3294
Y     130
Name: Risk_Rheumatoid_Arthritis, dtype: int64
-Column name is: Risk_Untreated_Chronic_Hyperthyroidism and it value is: N    3422
Y       2
Name: Risk_Untreated_Chronic_Hyperthyroidism, dtype: int64
-Column name is: Risk_Untreated_Chronic_Hypogonadism and it value is: N    3297
Y     127
Name: Risk_Untreated_Chronic_Hypogonadism, dtype: int64
-Column name is: Risk_Untreated_Early_Menopause and it value is: N    3412
Y      12
Name: Risk_Untreated_Early_Menopause, dtype: int64
-Column name is: Risk_Patient_Parent_Fractured_Their_Hip and it value is: N    3168
Y     256
Name: Risk_Patient_Parent_Fractured_Their_Hip, dtype: int64
-Column name is: Risk_Smoking_Tobacco and it value is: N    2780
Y     644
Name: Risk_Smoking_Tobacco, dtype: int64
-Column name is: Risk_Chronic_Malnutrition_Or_Malabsorption and it value is: N    2954
Y     470
Name: Risk_Chronic_Malnutrition_Or_Malabsorption, dtype: int64
-Column name is: Risk_Chronic_Liver_Disease and it value is: N    3406
Y      18
Name: Risk_Chronic_Liver_Disease, dtype: int64
-Column name is: Risk_Family_History_Of_Osteoporosis and it value is: N    3066
Y     358
Name: Risk_Family_History_Of_Osteoporosis, dtype: int64
-Column name is: Risk_Low_Calcium_Intake and it value is: N    3382
Y      42
Name: Risk_Low_Calcium_Intake, dtype: int64
-Column name is: Risk_Vitamin_D_Insufficiency and it value is: N    1788
Y    1636
Name: Risk_Vitamin_D_Insufficiency, dtype: int64
-Column name is: Risk_Poor_Health_Frailty and it value is: N    3232
Y     192
Name: Risk_Poor_Health_Frailty, dtype: int64
-Column name is: Risk_Excessive_Thinness and it value is: N    3357
Y      67
Name: Risk_Excessive_Thinness, dtype: int64
-Column name is: Risk_Hysterectomy_Oophorectomy and it value is: N    3370
Y      54
Name: Risk_Hysterectomy_Oophorectomy, dtype: int64
-Column name is: Risk_Estrogen_Deficiency and it value is: N    3413
Y      11
Name: Risk_Estrogen_Deficiency, dtype: int64
-Column name is: Risk_Immobilization and it value is: N    3410
Y      14
Name: Risk_Immobilization, dtype: int64
-Column name is: Risk_Recurring_Falls and it value is: N    3355
Y      69
Name: Risk_Recurring_Falls, dtype: int64
```

a. Handling outliers :

For numerical variables we have only 2 Dexa_Freq_During_Rx and Cuont_Of_Risks .

In order to detect if there is any outlier, we plot the boxplot:



We notice that we can not remove these outliers since we have a lot and the dataset is small, thus

We decided to use the winsorizing technique where any value of a variable above or below a percentile k on each side of the variables' distribution is replaced with the value of the k-th percentile itself. For example, if k=5, all observations above the 95th percentile are recoded to the value of the 95th percentile, and values below the 5th percent are recoded, respectively

```
from scipy.stats.mstats import winsorize
#Outer fences of the variable Dexa_Freq_During_Rx
def fences(df, variable_name):
    q1 = df[variable_name].quantile(0.25)
    q3 = df[variable_name].quantile(0.75)
    iqr = q3-q1
    outer_fence = 3*iqr
    outer_fence_le = q1-outer_fence
    outer_fence_ue = q3+outer_fence
    return outer_fence_le, outer_fence_ue
outer_fence_le, outer_fence_ue = fences(df, 'Dexa_Freq_During_Rx')
print('Lower end outer fence: ', outer_fence_le)
print('Upper end outer fence: ', outer_fence_ue)
```

```
Lower end outer fence:  -9.0
Upper end outer fence:  12.0
```

The upper outer fence for the variable "Dexa_Freq_During_Rx" is 12, while the lower end is below zero. Because a frequency below zero is not meaningful, the data should only be winsorized on its right tail. Now, we can look at values at different percentiles to set k.

```
print('90% quantile:   ', df['Dexa_Freq_During_Rx'].quantile(0.90))
print('92% quantile: ', df['Dexa_Freq_During_Rx'].quantile(0.92))
print('92.5% quantile: ', df['Dexa_Freq_During_Rx'].quantile(0.925))
print('95% quantile:   ', df['Dexa_Freq_During_Rx'].quantile(0.95))
print('97.5% quantile: ', df['Dexa_Freq_During_Rx'].quantile(0.975))
print('99% quantile:   ', df['Dexa_Freq_During_Rx'].quantile(0.99))
print('99.9% quantile: ', df['Dexa_Freq_During_Rx'].quantile(0.999))
```

```
90% quantile:    10.0
92% quantile:  11.0
92.5% quantile:  12.0
95% quantile:    14.0
97.5% quantile:  22.0
99% quantile:    34.76999999999998
99.9% quantile:  99.54000000000451
```

At 92.5% (12) we reach the upper outer fence. Hence i will winsorize the data on k=7.5 using the winsorize function from scipy:

```
#Create copy of df
df_win = df.copy(deep=True)

#Winsorize on right-tail
df_win['Dexa_Freq_During_Rx_wins_925%'] = winsorize(df['Dexa_Freq_During_Rx'], limits=(0, 0.075))

print(df_win.describe())
plt.figure(figsize=[10, 8])

sns.distplot(df['Dexa_Freq_During_Rx'])

#New distribution plots
sns.distplot(df_win['Dexa_Freq_During_Rx_wins_925%'])
plt.show()
```

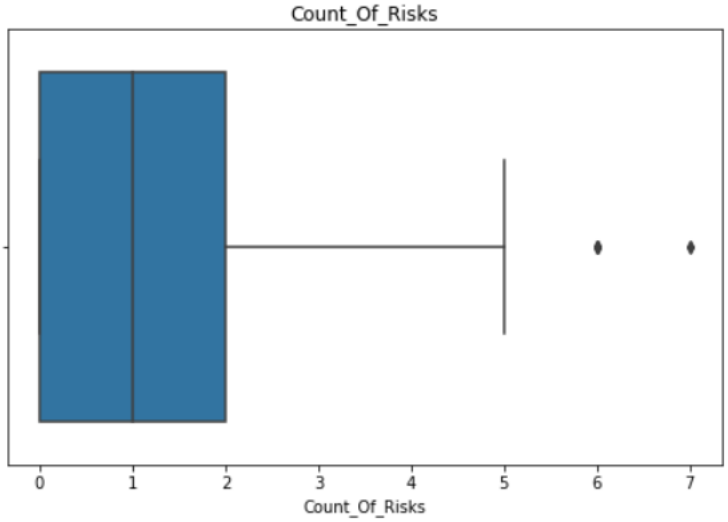
```
       Dexa_Freq_During_Rx  Count_Of_Risks  Dexa_Freq_During_Rx_wins_925%
count          3424.000000     3424.000000                    3424.000000
mean              3.016063        1.239486                       2.145152
std               8.136545        1.094914                       3.918497
min               0.000000        0.000000                       0.000000
25%               0.000000        0.000000                       0.000000
50%               0.000000        1.000000                       0.000000
75%               3.000000        2.000000                       3.000000
max             146.000000        7.000000                      12.000000
```

Now we can see the difference after winsorizing the max is 12 instead of 146

```
df_win['Dexa_Freq_During_Rx_wins_925%'].value_counts()
```

```
0     2488
12     272
5      114
6      107
7       93
8       71
4       68
10      55
3       46
9       32
11      30
2       24
1       24
Name: Dexa_Freq_During_Rx_wins_925%, dtype: int64
```

For Count_Of_Risks variable :



We notice that we don't have lot of outliers only 6 and 7 values, now let's check the occurrence of each of them to decide which technique will be good for handling these points.

```
df['Count_Of_Risks'].value_counts()
```

```
1    1242
0     970
2     781
3     317
4      91
5      15
6       6
7       2
Name: Count_Of_Risks, dtype: int64
```

We notice that in total we have only 8 values considered as outliers so we can remove them , it won't cause a problem:

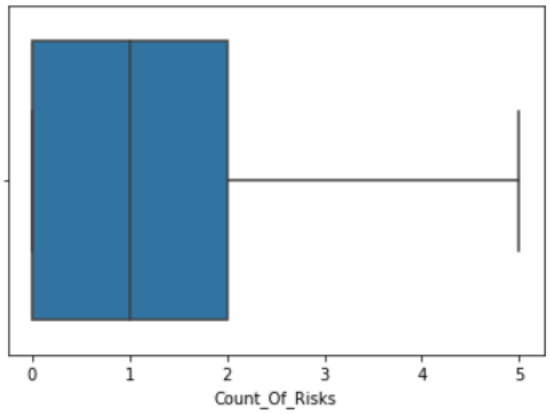
```
df = df[(df['Count_Of_Risks']<6)]
df['Count_Of_Risks'].value_counts()
```

```
1    1242
0     970
2     781
3     317
4      91
5      15
Name: Count_Of_Risks, dtype: int64
```

After selecting only the rows that have values lower than 6, this is how the boxplot looks like this time:

```
sns.boxplot(df['Count_Of_Risks'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe44147fd10>
```



Now since we're done with nan values and outliers…our dataset is ready for EDA from where we will get more insights and information about the data.