Healthcare Persistency of a drug

Data Science project

August 2021

COOL DATA SCIENTISTS TEAM

Name	Email	Country	College / Company	Specialization
Yousef Elbayoumi	yousefxelbayomi@gmail.com	Palestine	Bahçeşehir University	Data Science
Mukhammadjon Kholmirzev	kmukhammadjon@gmail.com	Uzbekistan	Ulsan National Institute of Science and Technology	Data Science
Jamila hamdi	jamila.hamdi90@gmail.com	Tunisia	Trainee	Data Science
H. Melis Tekin Akcin	meliss85@gmail.com	UK	Hacettepe University	Data Science

Contents

- Background
- Statistical Analysis
- EDA performed on the data
- Final Recomendations

Background

Problem Statement:

- One of the challenges for Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. This issue results in a bad impact on the pharmacies for all the categories; patients, physicians, and administration. However, the team of data scientist is capable of discovering the analyzing the dataset and detecting the factors that are impacting the primary factor which is the "persistency". By building a classification machine learning model, we will be able to classify the dataset and find the variables that affect the target variables "Persistency Flag".

EDA performed on data

Dataset:

```
df.head()
```

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...
0	P1	Persistent	Male	Caucasian	Not Hispanic	West	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West	55-65	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest	65-75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...

Totally we have 3424 observations and 69 features.

EDA performed on data

Dataset:

```
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
```

EDA performed on data

Features types:

```
df.dtypes
```

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

Null Values

```
In [8]: df.isnull().values.any()

Out[8]: False

In [9]: df.isnull().sum()

Out[9]: 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
```

We checked from the data and didn't find any null values.

Unknown Values

```
In [9]: df["Ethnicity"].value_counts()

Out[9]: Not Hispanic    3235
        Hispanic          98
        Unknown           91
        Name: Ethnicity, dtype: int64
```

```
In [11]: df["Region"].value_counts()

Out[11]: Midwest          1383
         South            1247
         West              502
         Northeast         232
         Other/Unknown      60
         Name: Region, dtype: int64
```

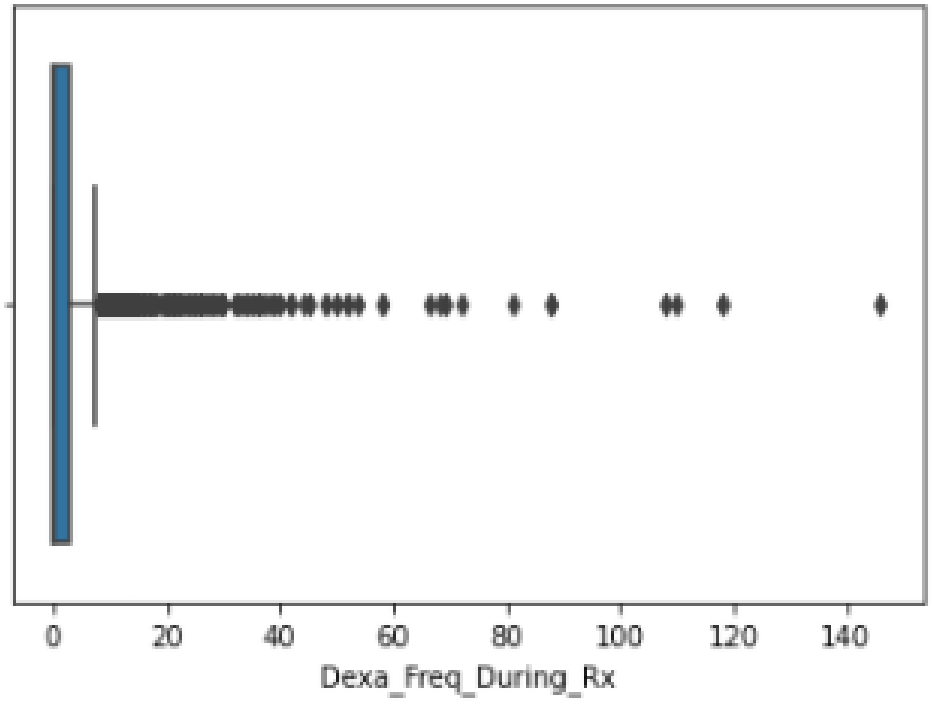
```
In [20]: df["Risk_Segment_During_Rx"].value_counts()

Out[20]: Unknown    1497
         HR_VHR      965
         VLR_LR      962
         Name: Risk_Segment_During_Rx, dtype: int64
```

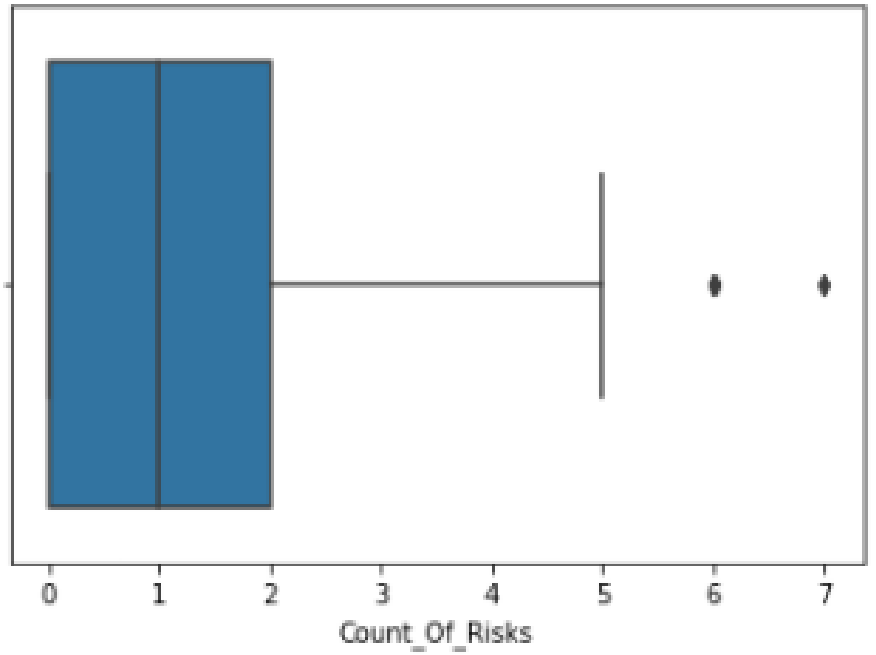
On the other hand, we found a lot of the "Unknown" values, we considered them as null values and decided to remove them because they can affect the results of our ML models.

Outliers

We have 460 outliers in "Dexa_Freq_During_Rx" variable.



We have 8 outliers in "Count_Of_Risks" variable.

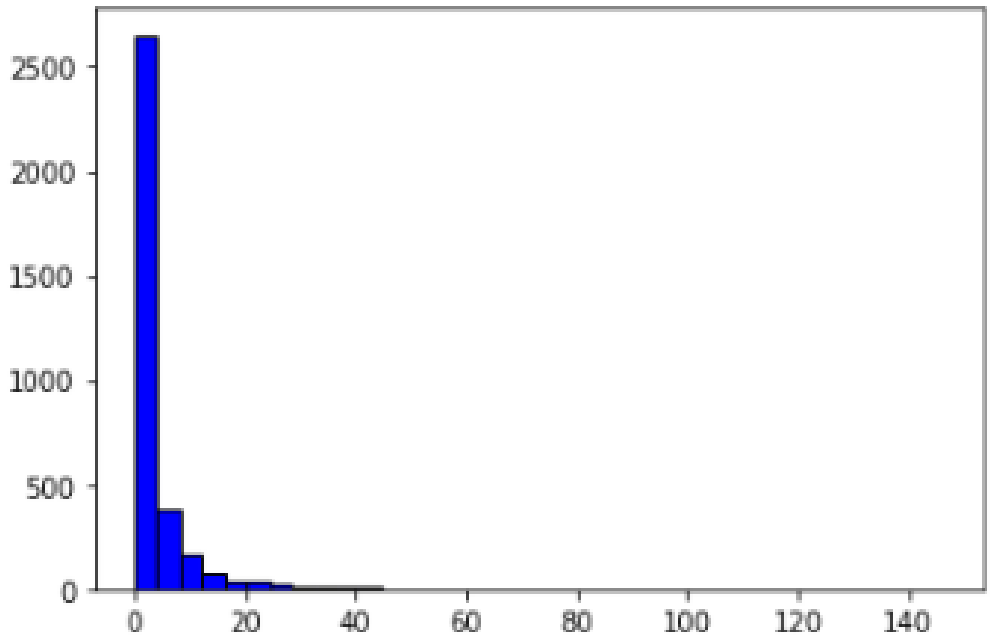


Skewed Data

As seen here, since the tail is on the right side, we can say that "Dexa_Freq_During_Rx" variable has **right-skewed distribution**.

Hence, we can conclude that the mean value is greater than the mode.

histogram graph:



EDA performed on data

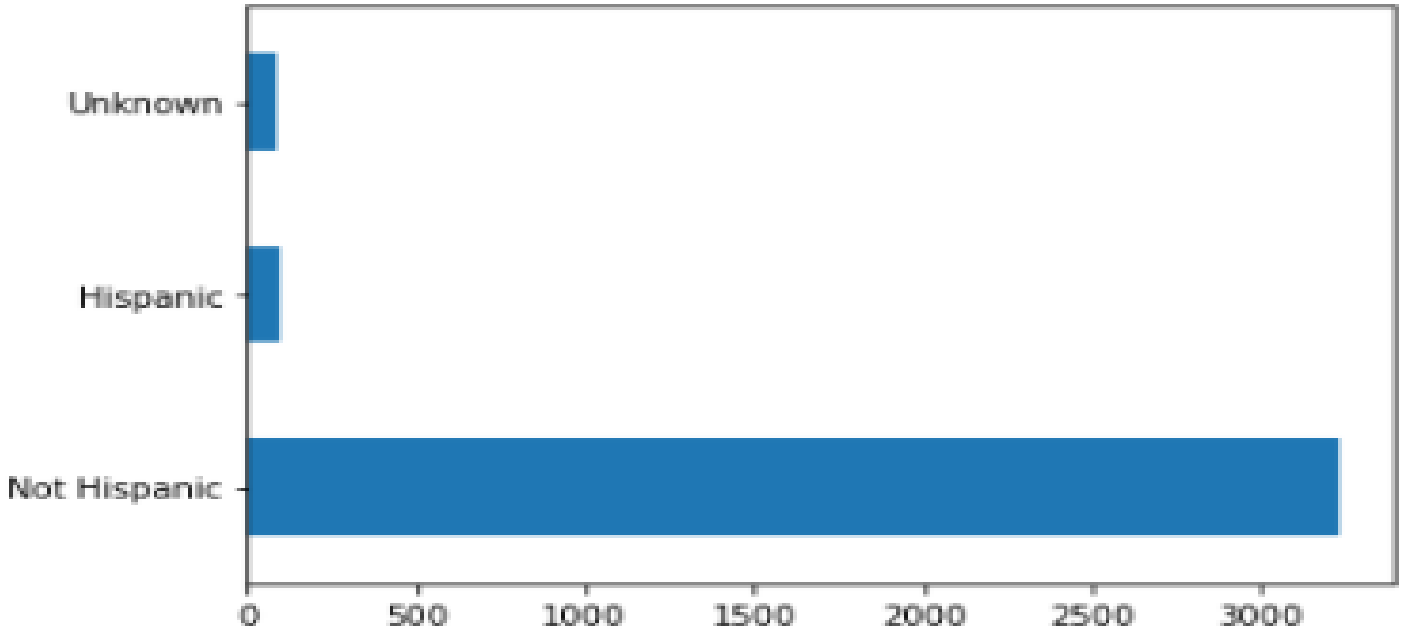
Features types:

```
df.dtypes
```

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

EDA performed on data

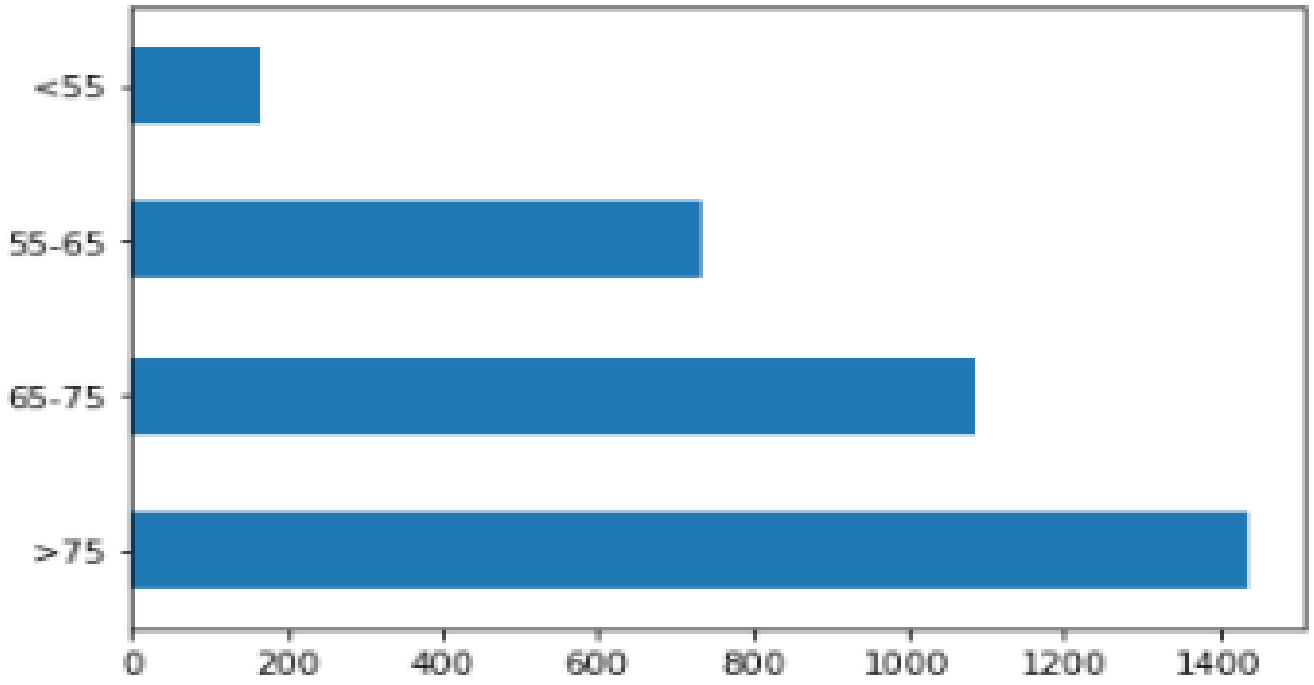
Demoghraphics analysis:



Ethnicity : "Non-Hispanic" people dominates the "Hispanic" people and also we have unknown values.

EDA performed on data

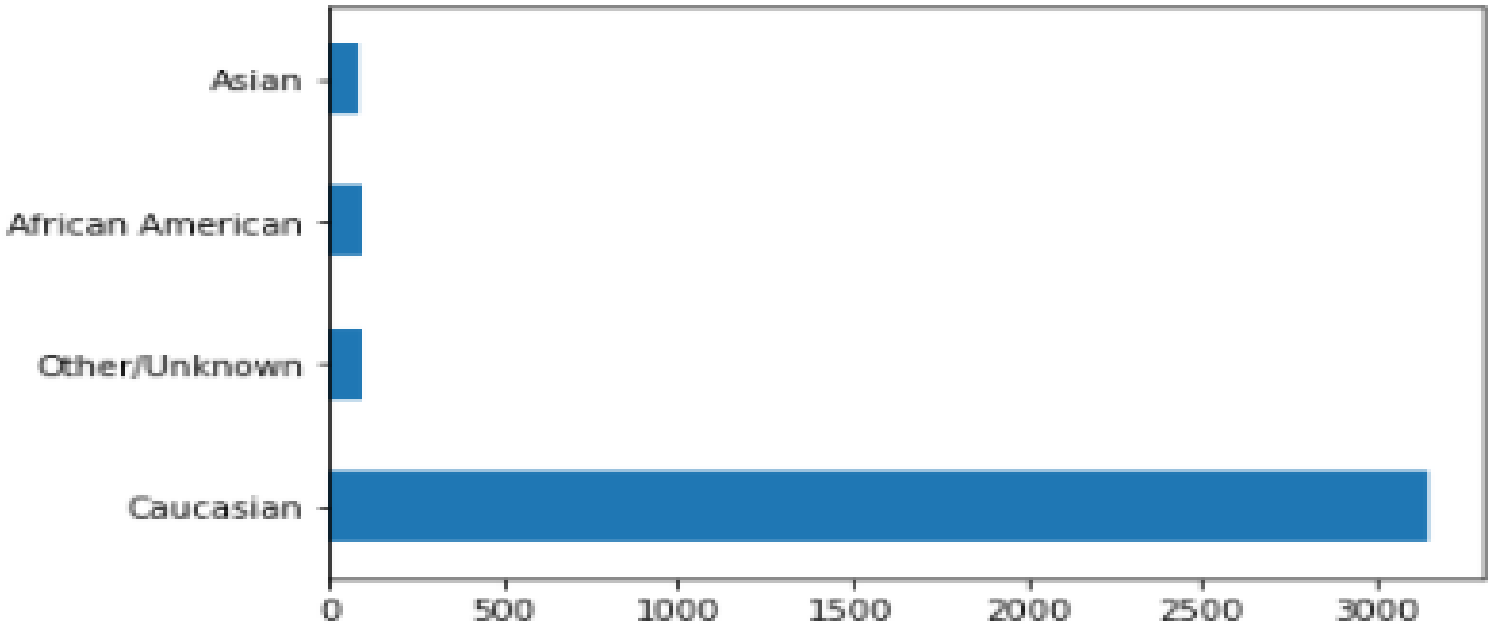
Demoghraphics analysis:



Age: age “>55” can be related to have persistency to drug.

EDA performed on data

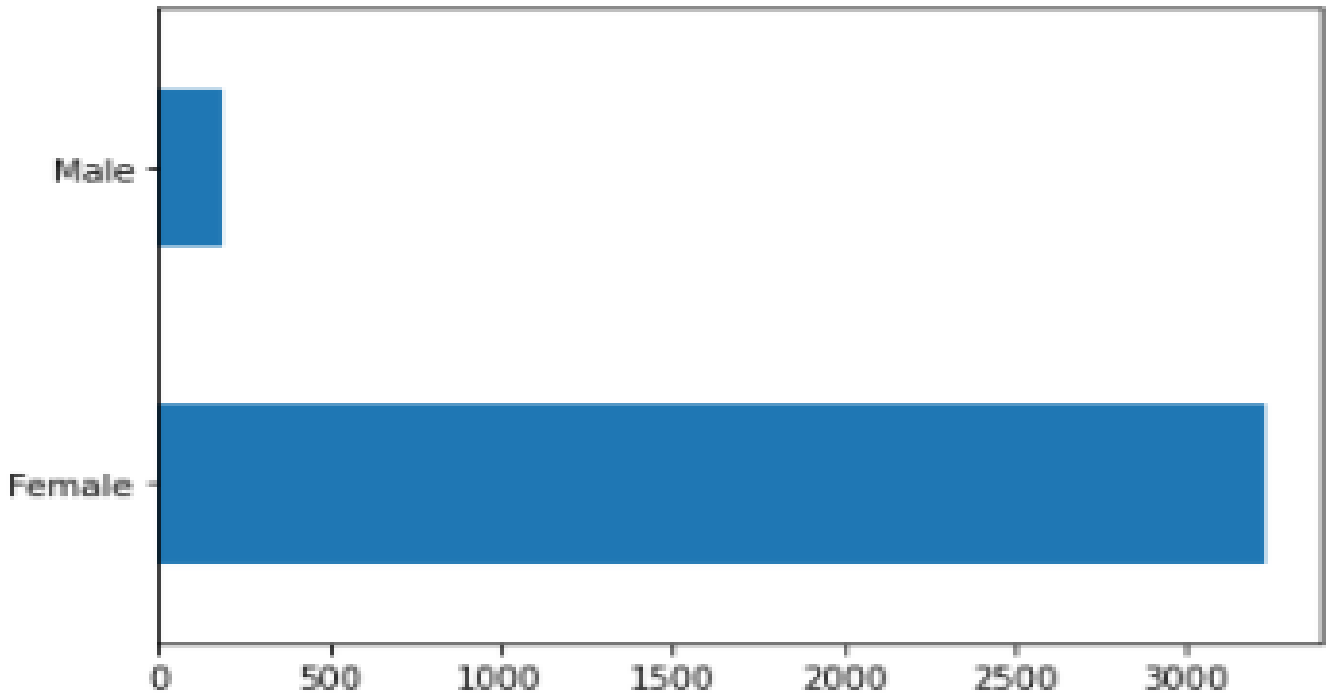
Demoghraphics analysis:



Race: the Caucasians are dominated the other races.

EDA performed on data

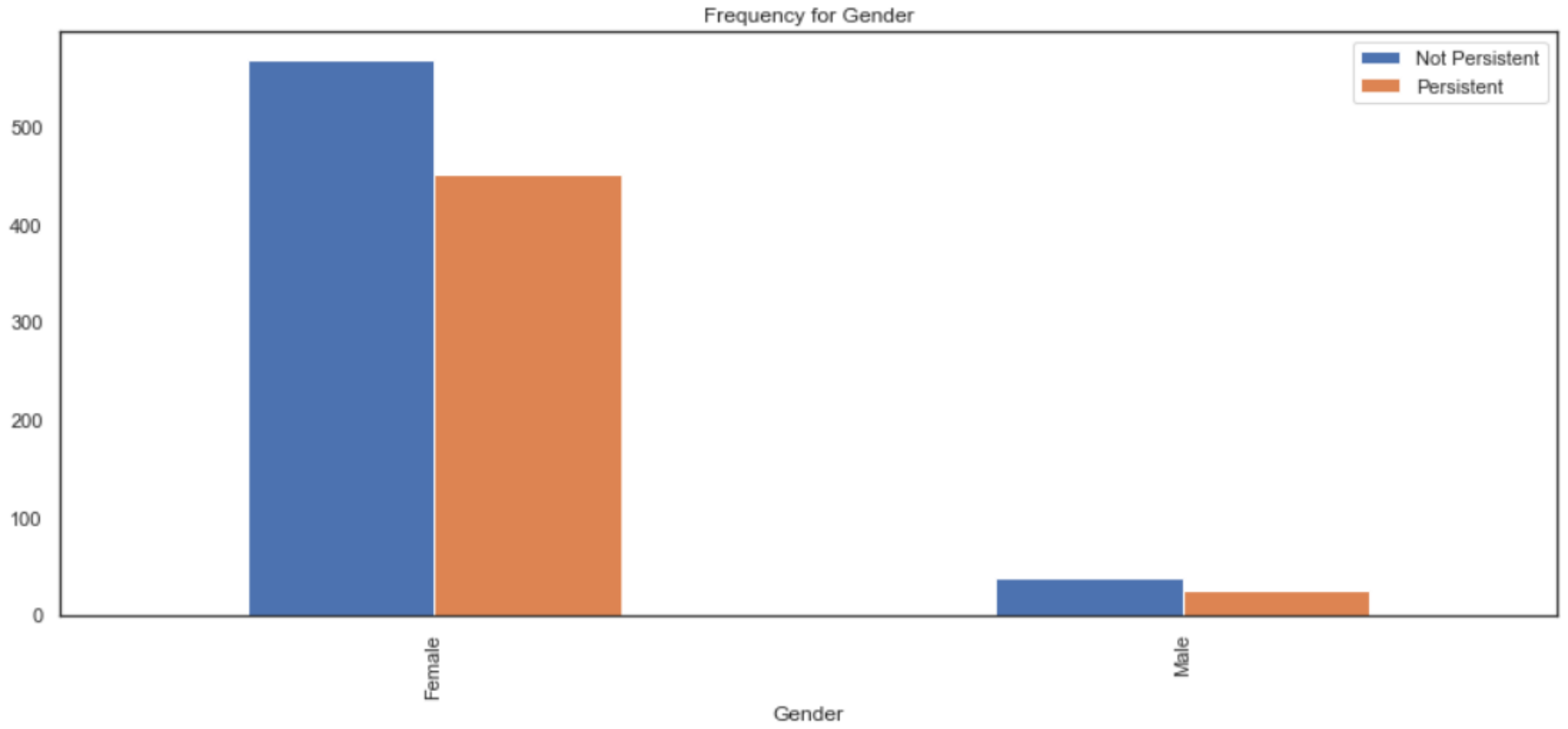
Demoghraphics analysis:



Gender: the female patients are more than the male patients.

EDA performed on data

Demoghraphics analysis:



Gender wise Analysis : As you can see from the graph, a huge imbalance between the genders

EDA performed on data

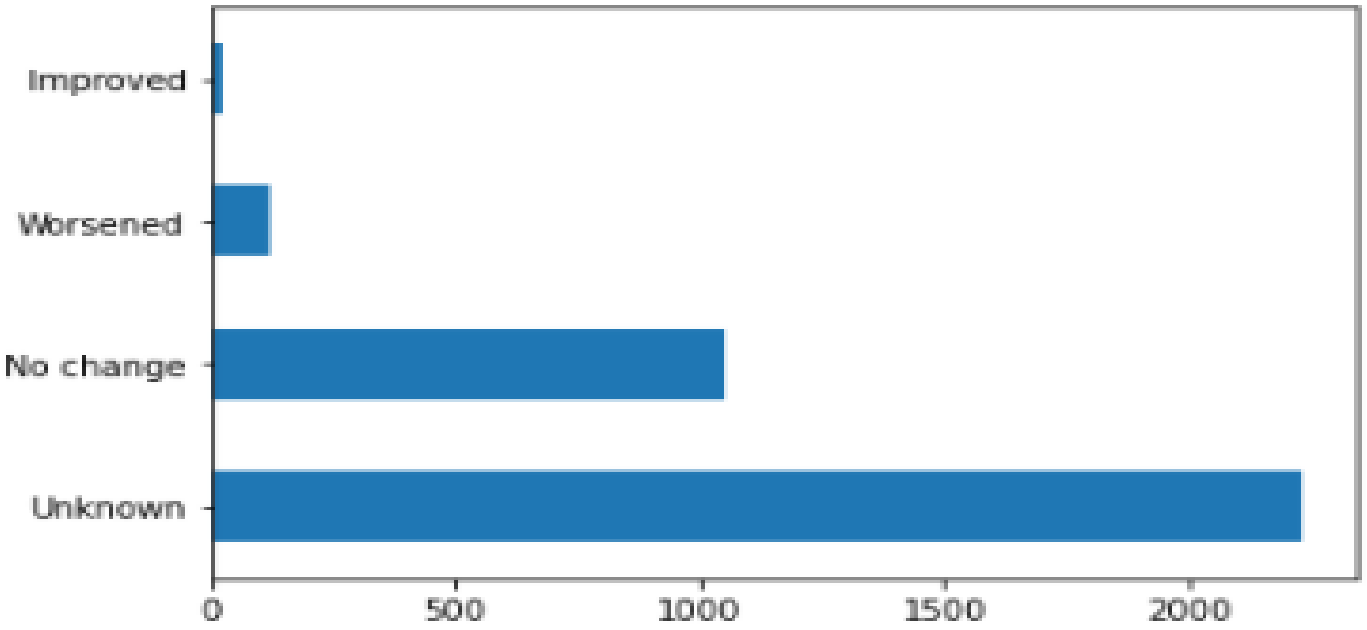
Ntm Speciality analysis:

General Practitioner, Rheumatology, Endocrinology and Oncology specialists prescribed the NTM Rx most.

GENERAL PRACTITIONER	1535
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
PSYCHIATRY AND NEUROLOGY	4
NEPHROLOGY	3
ORTHOPEDICS	3
GERIATRIC MEDICINE	2
HOSPICE AND PALLIATIVE MEDICINE	2
PLASTIC SURGERY	2
GASTROENTEROLOGY	2
VASCULAR SURGERY	2
TRANSPLANT SURGERY	2
OCCUPATIONAL MEDICINE	1
OPHTHALMOLOGY	1
PAIN MEDICINE	1

EDA performed on data

Clinical Factors analysis:



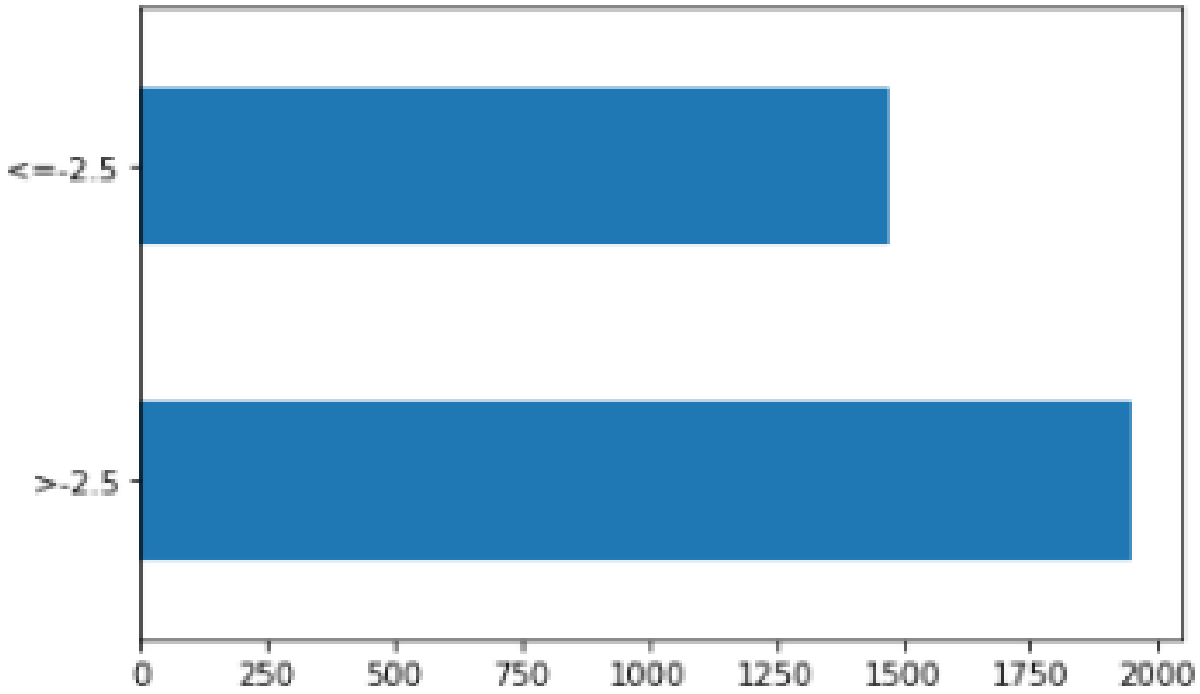
Fragility: we have obtained the following cross- table:

Frag_Frac_During_Rx	N	Y
Frag_Frac_Prior_Ntm		
N	2691	181
Y	316	236

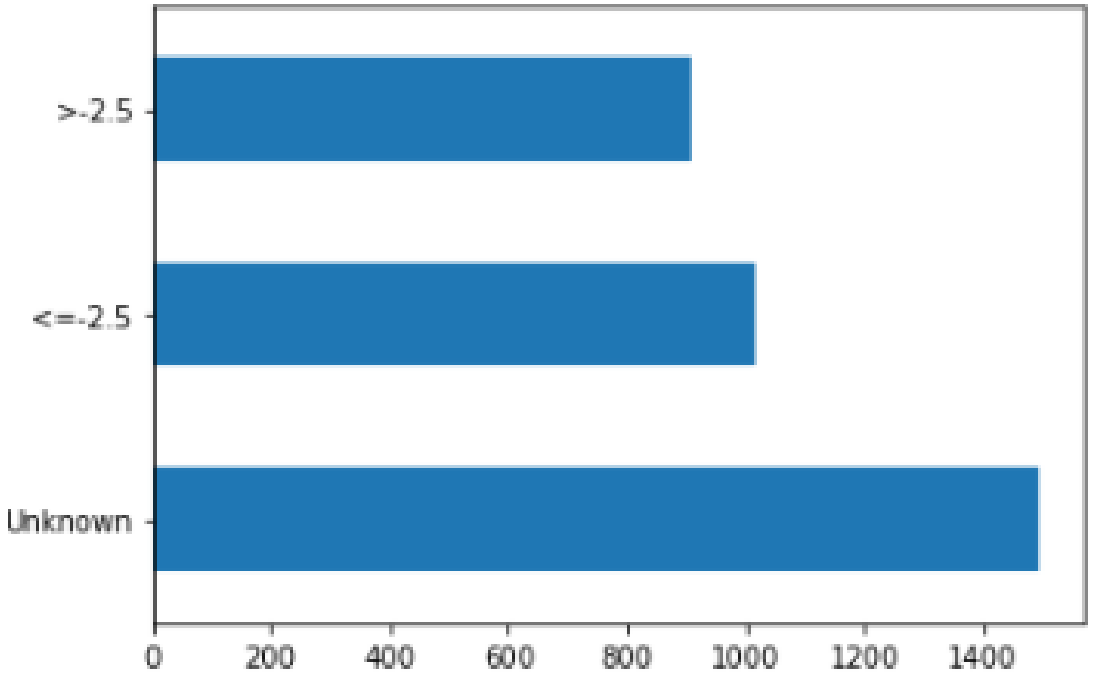
Risk Segment: We have compared the risk segments prior NTM and during NTM and examine how it changes:

EDA performed on data

T-scores prior to NTM:



T-scores during RX:



EDA performed on data

Statistics analysis:

	count	mean	std	min	25%	50%	75%	max
Persistency_Flag	1081.0	0.439408	0.496545	0.0	0.0	0.0	1.0	1.0
Gender	1081.0	0.058279	0.234379	0.0	0.0	0.0	0.0	1.0
Race	1081.0	1.916744	0.435153	0.0	2.0	2.0	2.0	3.0
Ethnicity	1081.0	0.966698	0.179508	0.0	1.0	1.0	1.0	1.0
Region	1081.0	1.832562	1.622953	0.0	0.0	3.0	3.0	4.0
...	...	...	...	...	...	...	...	...
Risk_Hysterectomy_Oophorectomy	1081.0	0.016651	0.128020	0.0	0.0	0.0	0.0	1.0
Risk_Estrogen_Deficiency	1081.0	0.000925	0.030415	0.0	0.0	0.0	0.0	1.0
Risk_Immobilization	1081.0	0.002775	0.052631	0.0	0.0	0.0	0.0	1.0
Risk_Recurring_Falls	1081.0	0.029602	0.169566	0.0	0.0	0.0	0.0	1.0
Count_Of_Risks	1081.0	1.457909	1.118173	0.0	1.0	1.0	2.0	7.0

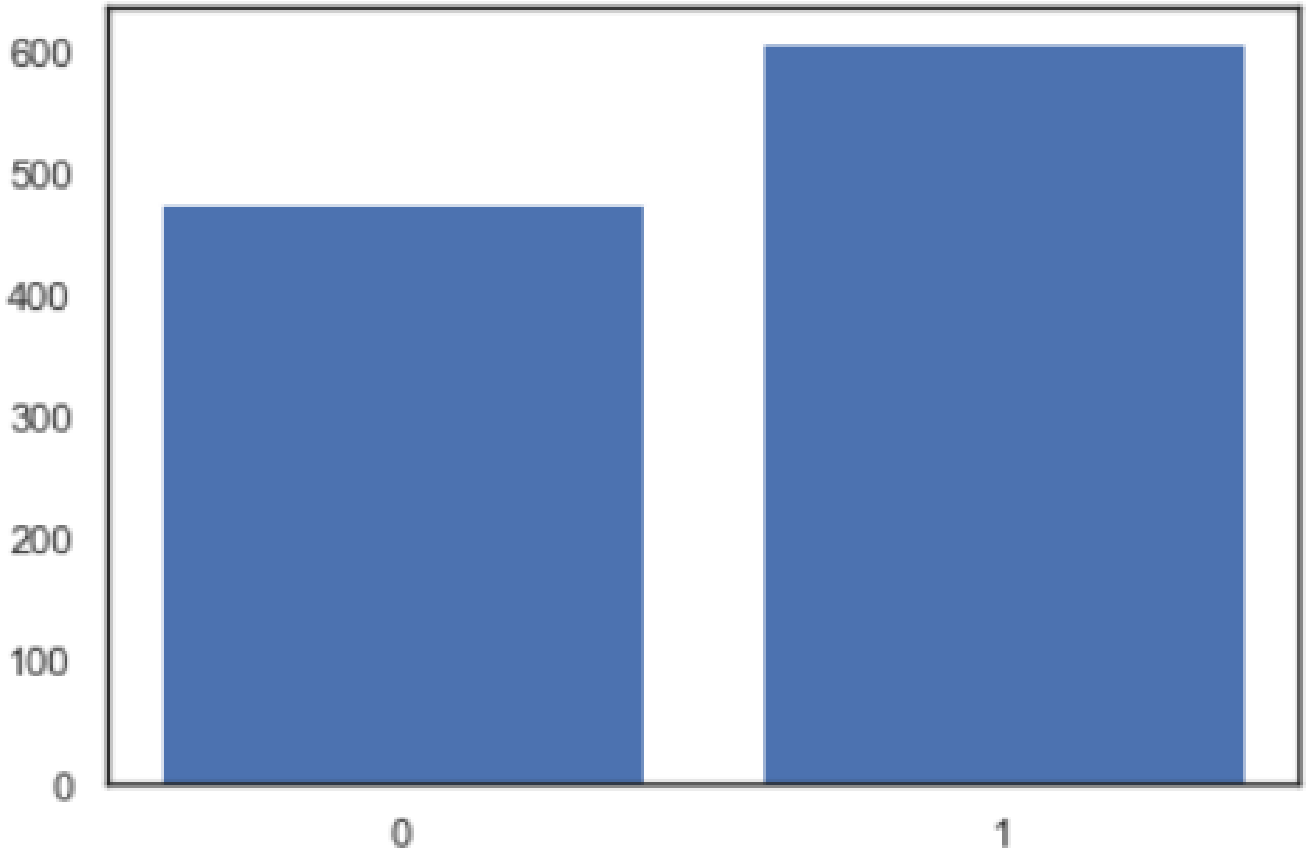
Statistics for numerical Features

	count	unique	top	freq
Ptid	1081	1081	P552	1
Risk_Segment_During_Rx	1081	2	HR_VHR	827
Tscore_Bucket_During_Rx	1081	2	<=-2.5	779
Change_T_Score	1081	3	No change	962
Change_Risk_Segment	1081	3	No change	953

Statistics for categorical features

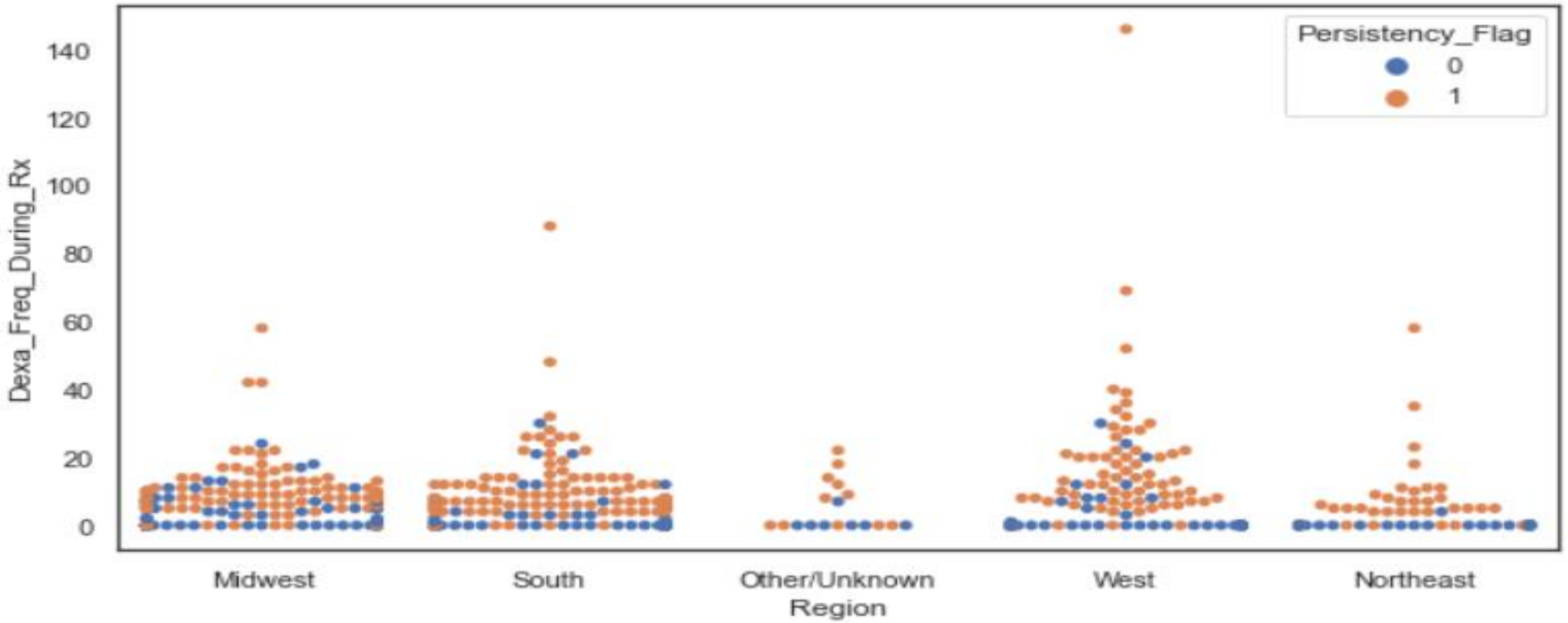
EDA performed on data

Ratio of the target variable:

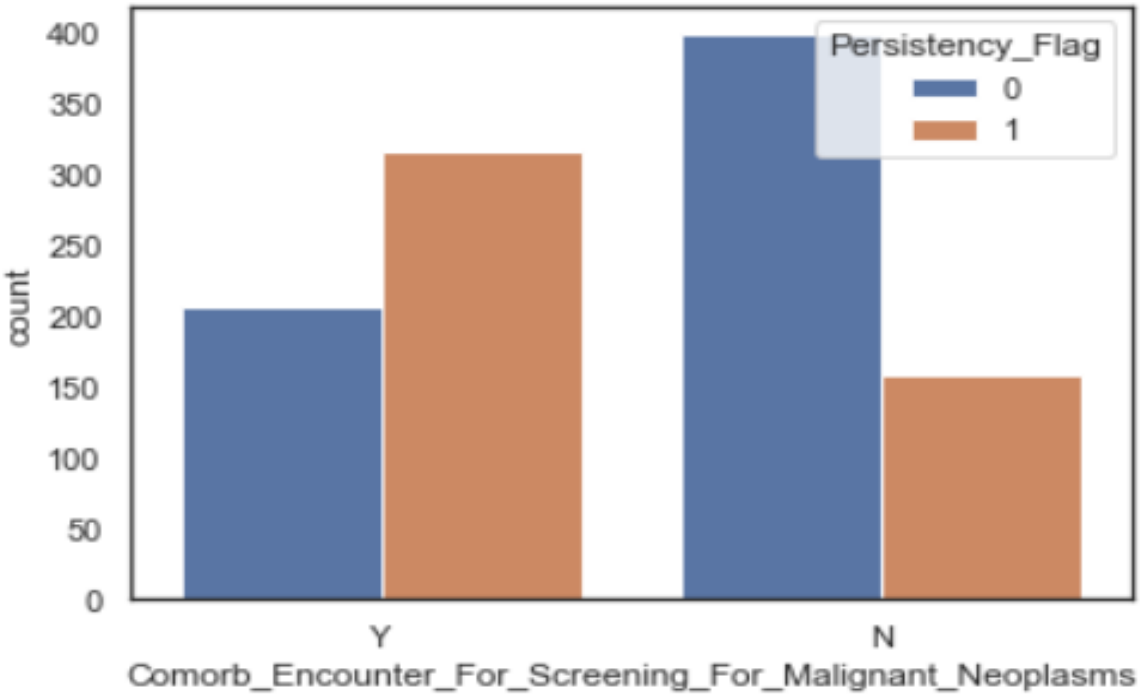


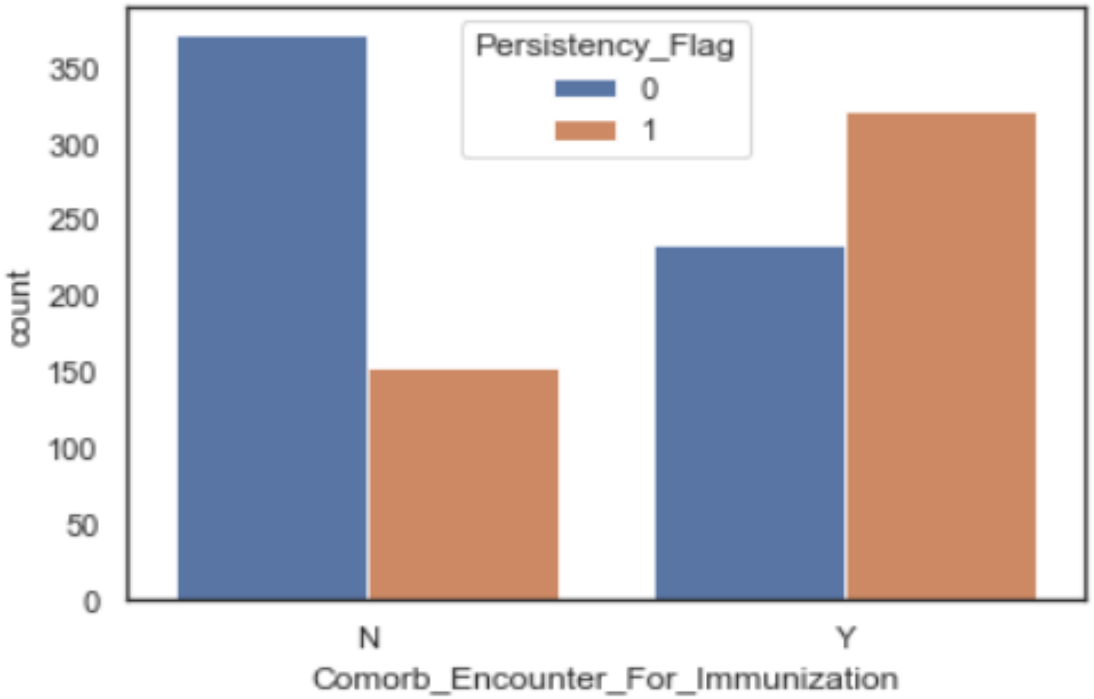
EDA performed on data

Number of DEXA scans by each region

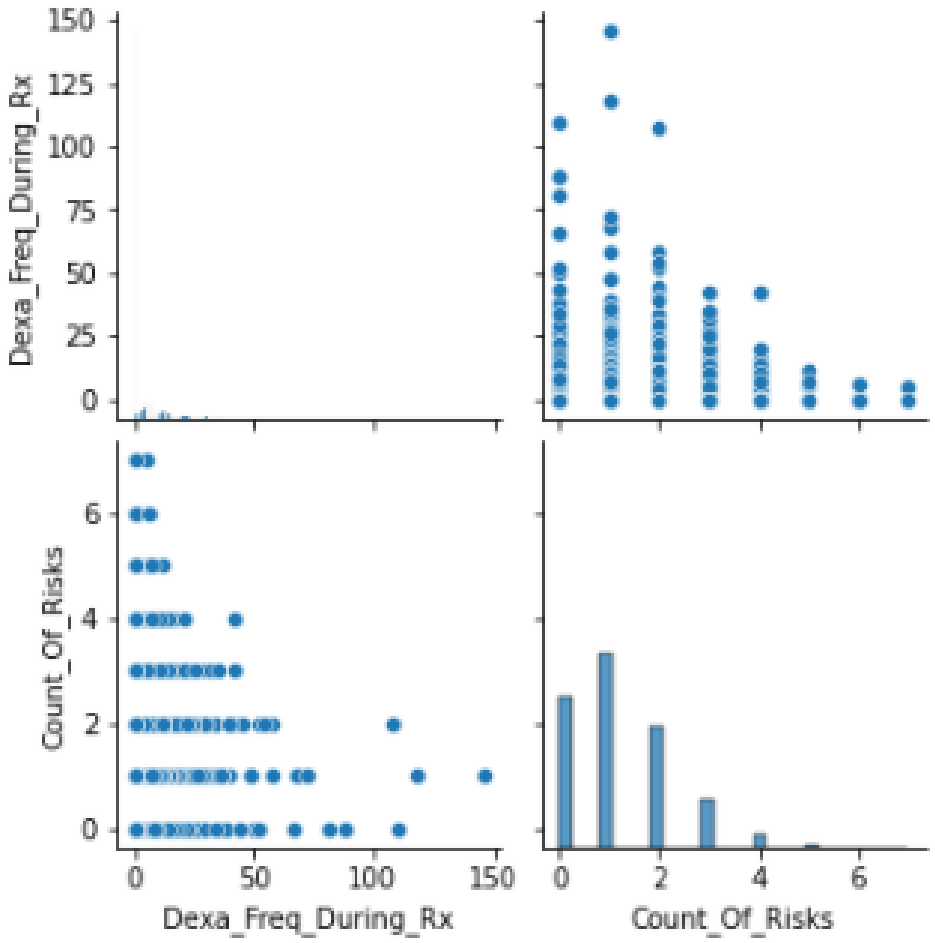


EDA performed on data





Numerical Values



We have only two columns with numerical values, these diagrams shows the relations between these columns.

Final recommendations

- According to Cleaning: The data is considered clean pretty much.
- According to Region: The data mostly as “Not Persistent”.
- According to Correlations: The data doesn’t have good correlations due to encoding the data, we replaced Y and N with 0 and 1.
- According to Statistics: Similar to the correlations, we can’t comment much here due to the same reason.
- Obviously, this is a classification problem, the team is considering several ML models to build and test, such as KNN, MLP, Decision Tree, Random Forest, etc.

Thank You