Data Glacier

Your Deep Learning Partner

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

Data Glacier
Your Deep Learning Partner

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".

Statistical Analysis

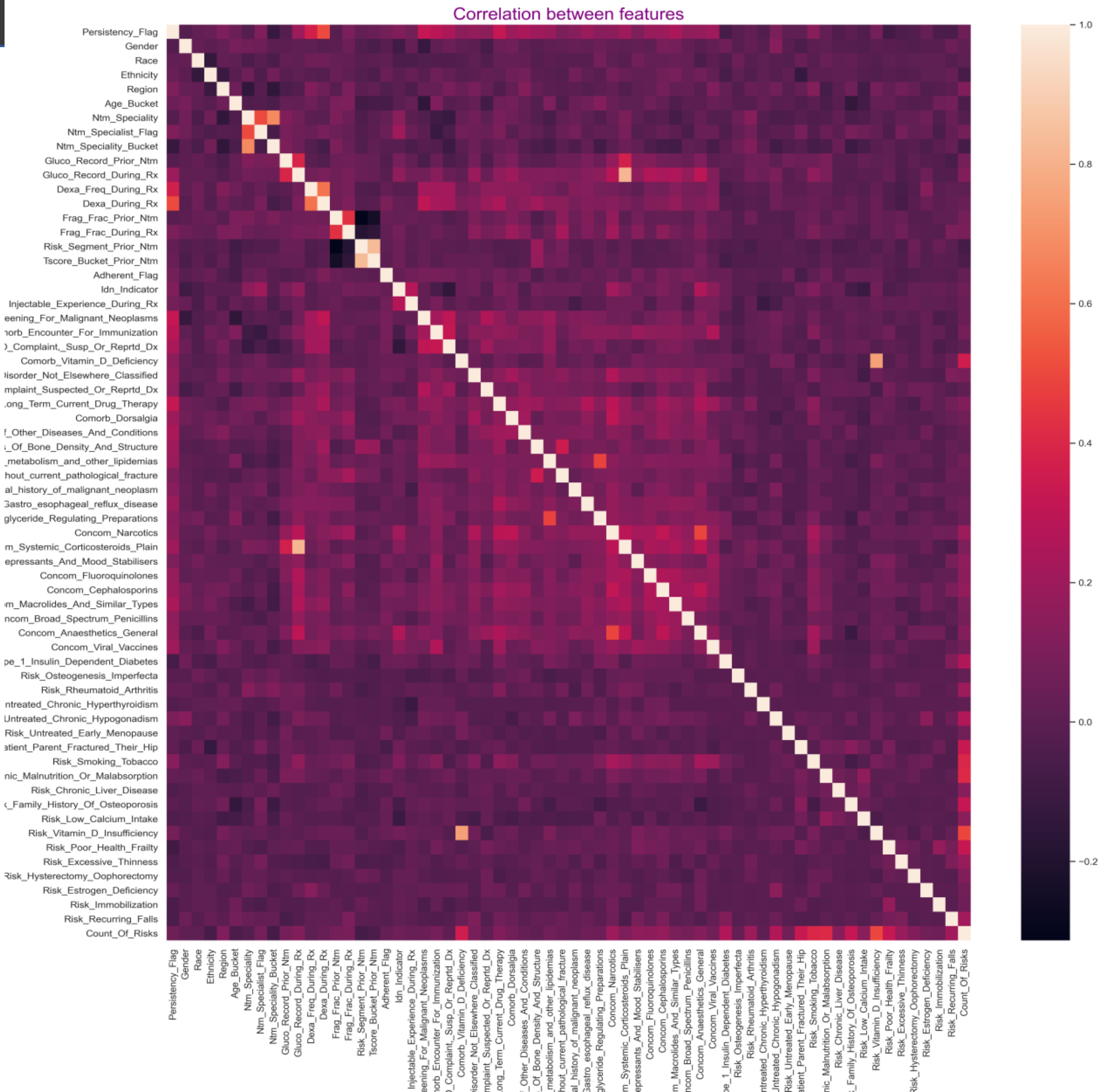
	count	mean	std	min	25%	50%	75%	max
Persistency_Flag	1081.0	0.439408	0.496545	0.0	0.0	0.0	1.0	1.0
Gender	1081.0	0.058279	0.234379	0.0	0.0	0.0	0.0	1.0
Race	1081.0	1.916744	0.435153	0.0	2.0	2.0	2.0	3.0
Ethnicity	1081.0	0.966698	0.179508	0.0	1.0	1.0	1.0	1.0
Region	1081.0	1.832562	1.622953	0.0	0.0	3.0	3.0	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

Correlation Analysis

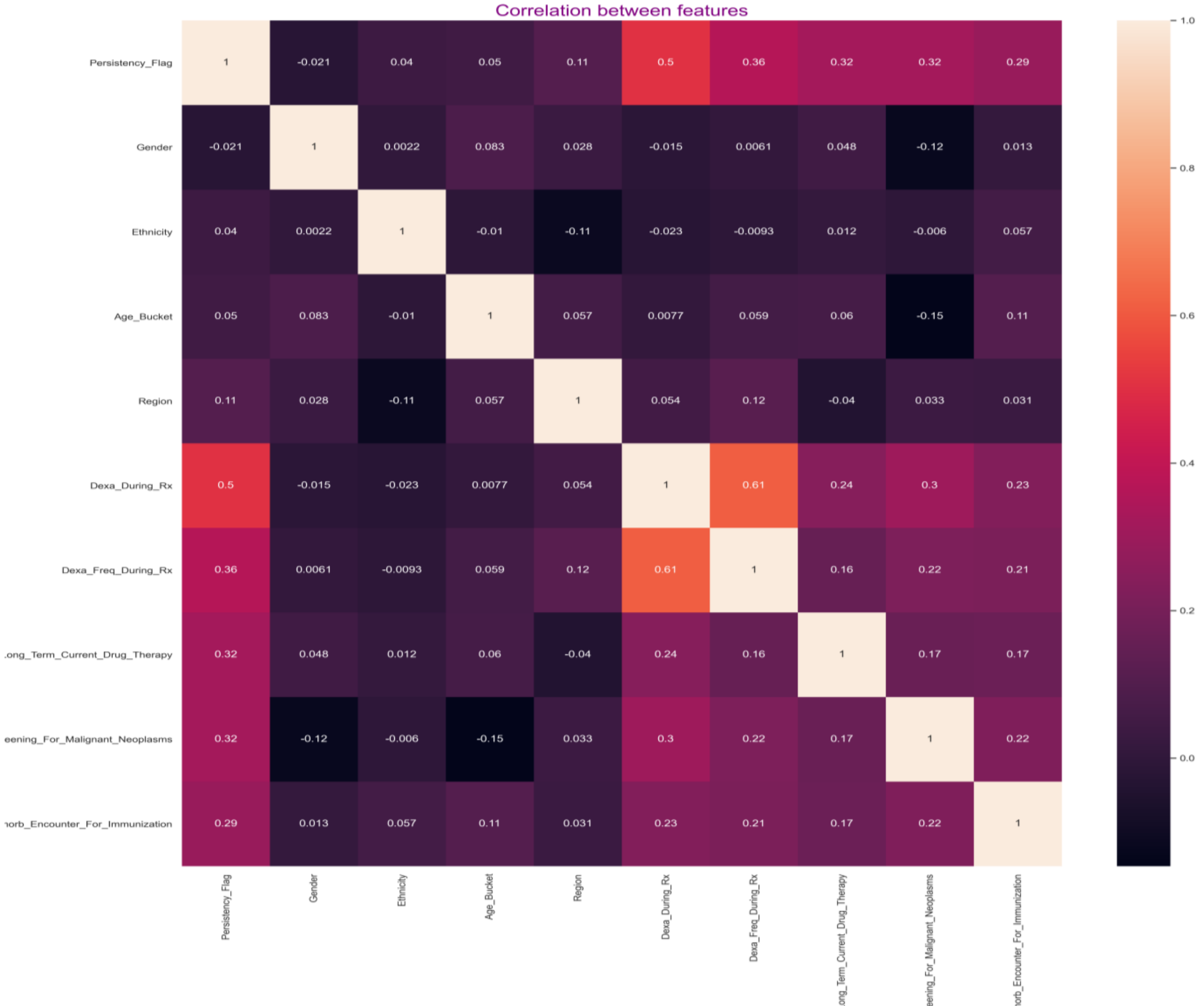


Correlation Analysis

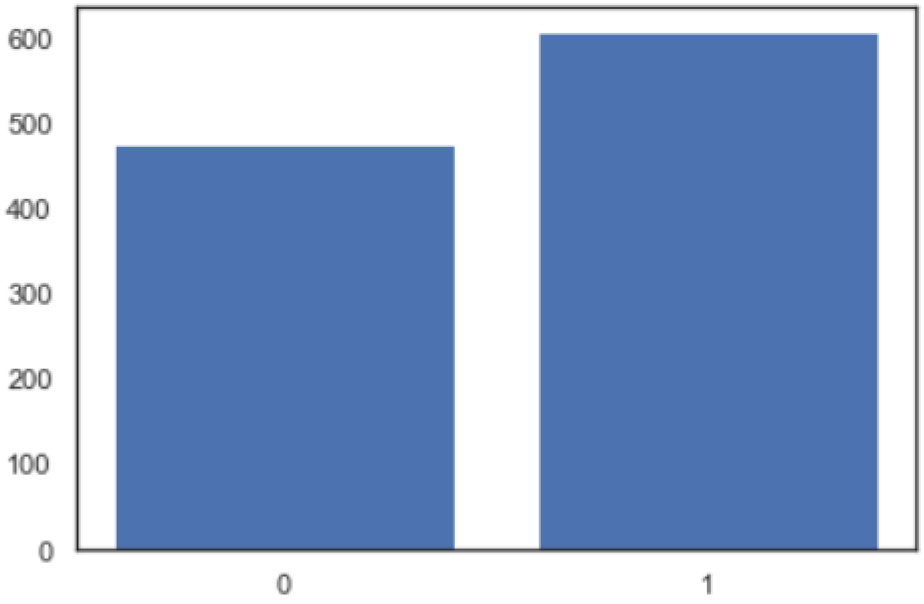
Correlation of features to the target

```
Persistency_Flag                                          1.000000
Dexa_During_Rx                                            0.503839
Dexa_Freq_During_Rx                                       0.364767
Comorb_Long_Term_Current_Drug_Therapy                     0.323256
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms    0.321462
Comorb_Encounter_For_Immunization                         0.289725
Comorb_Personal_History_Of_Other_Diseases_And_Conditions  0.244608
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx  0.243454
Concom_Macrolides_And_Similar_Types                       0.239266
Comorb_Other_Disorders_Of_Bone_Density_And_Structure      0.208364
Name: Persistency_Flag, dtype: float64
```

Top 10 most correlated features to the target



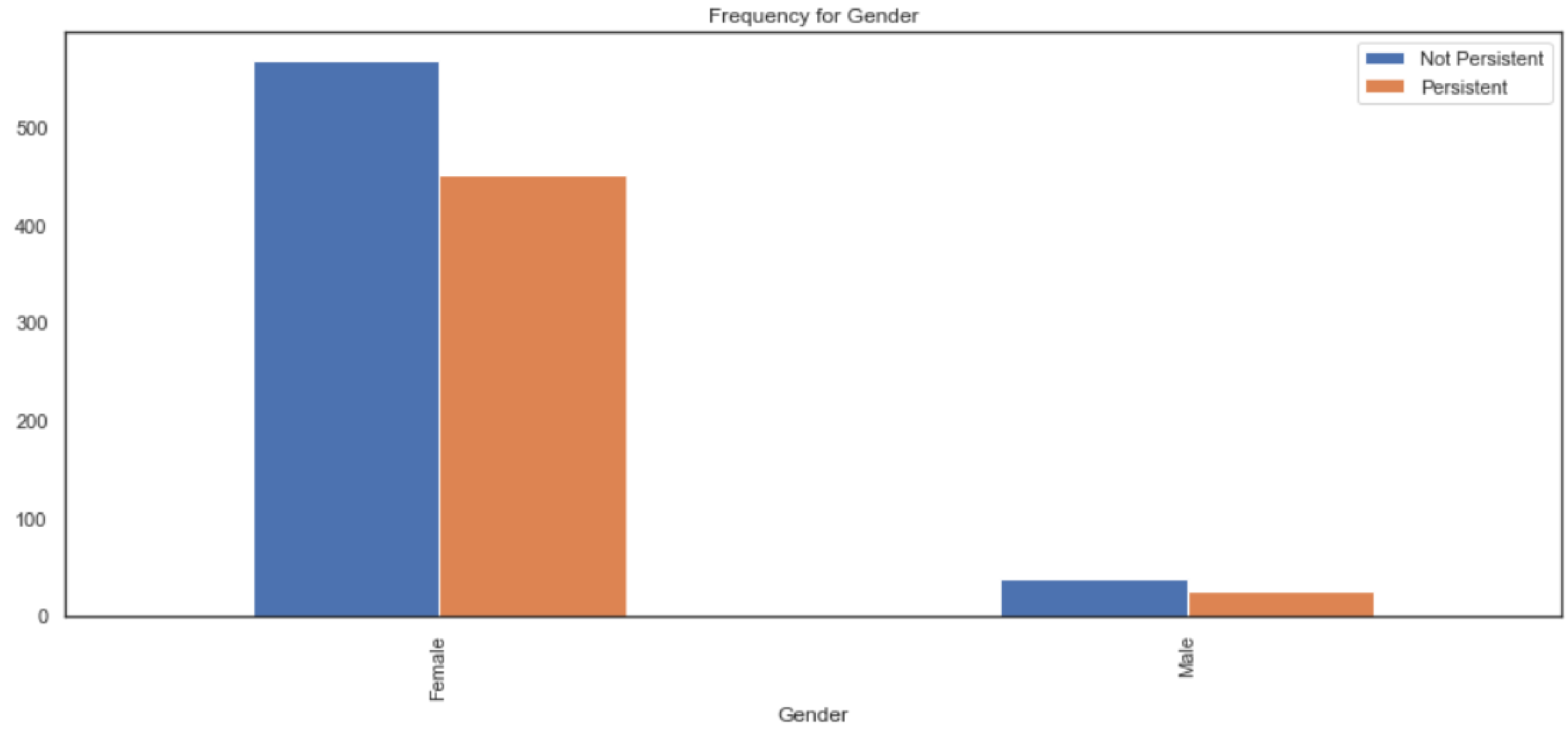
Hypotethis & Data Visualization



Checking the ratio of the target variable

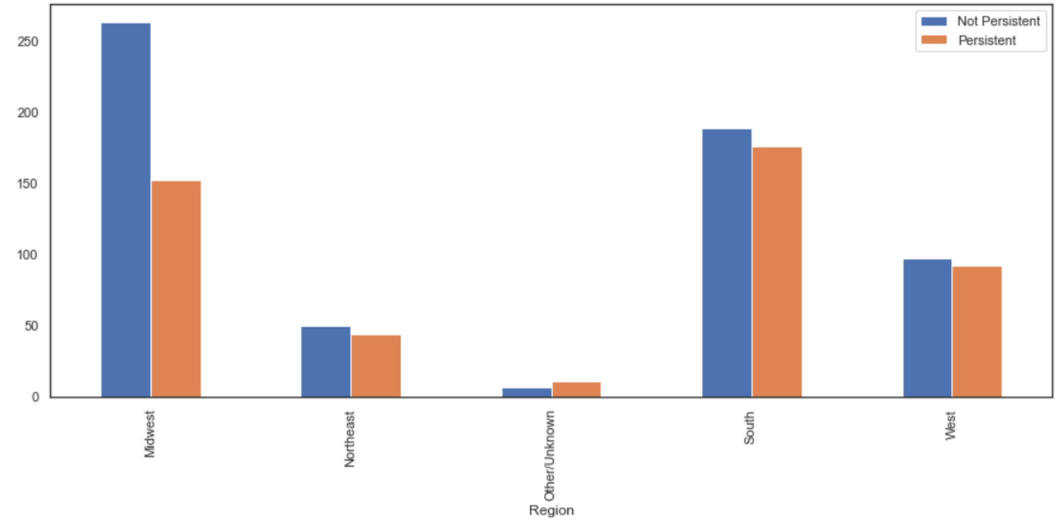
Hypotethis & Data Visualization

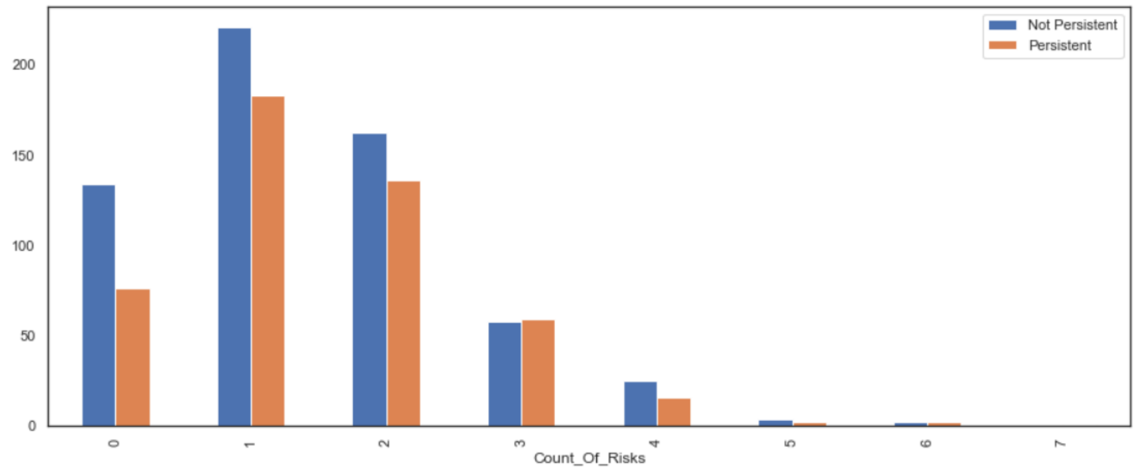
Gender wise Analysis



As you can see from the graph, a huge imbalance between the genders

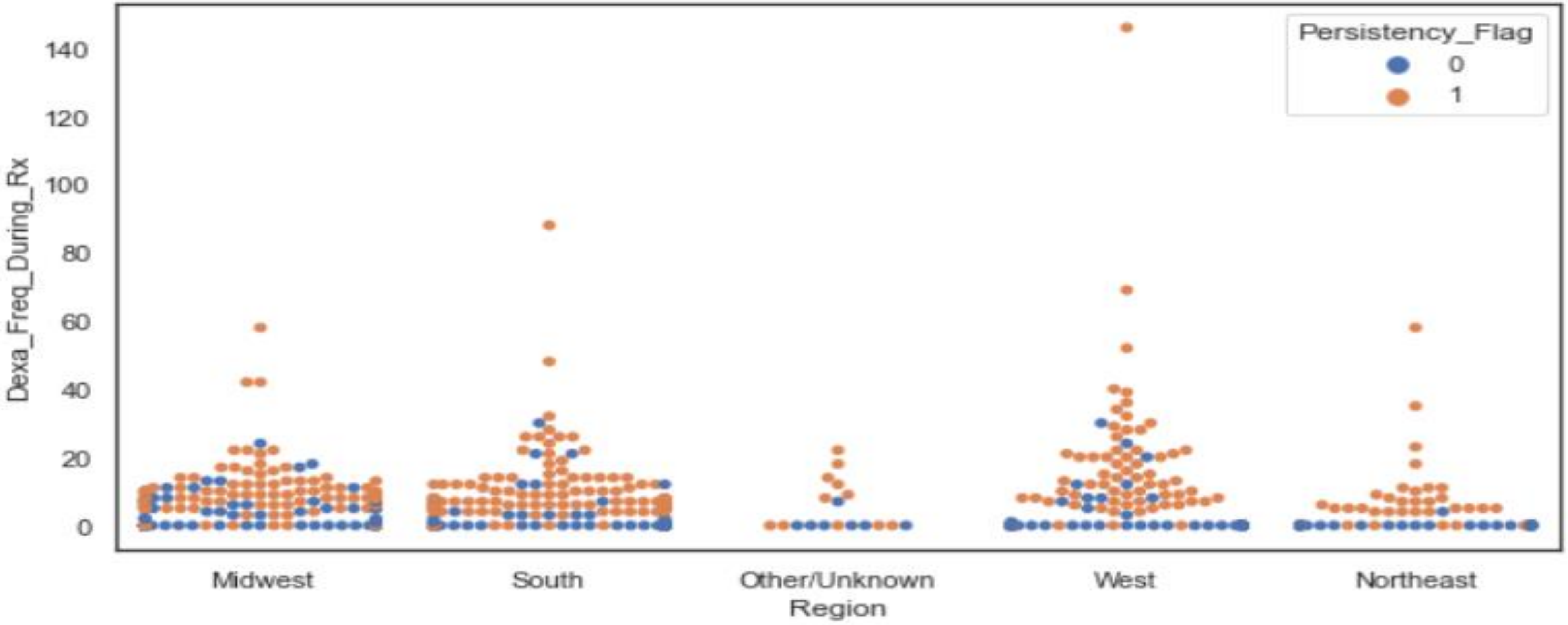
Hypotethis & Data Visualization



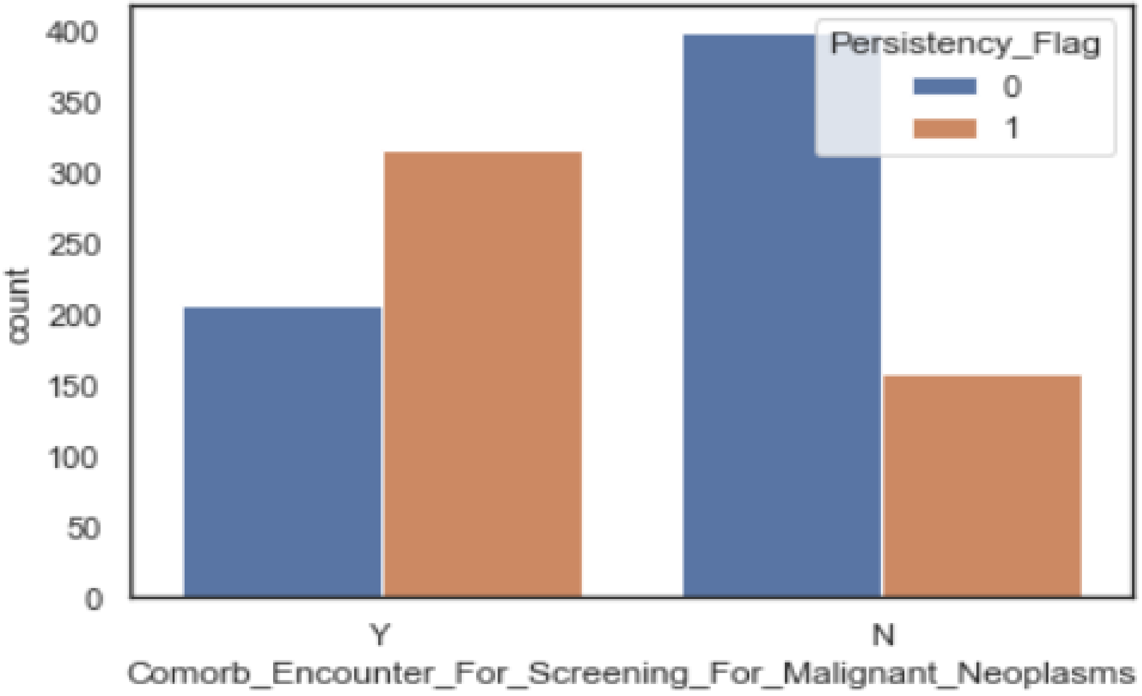


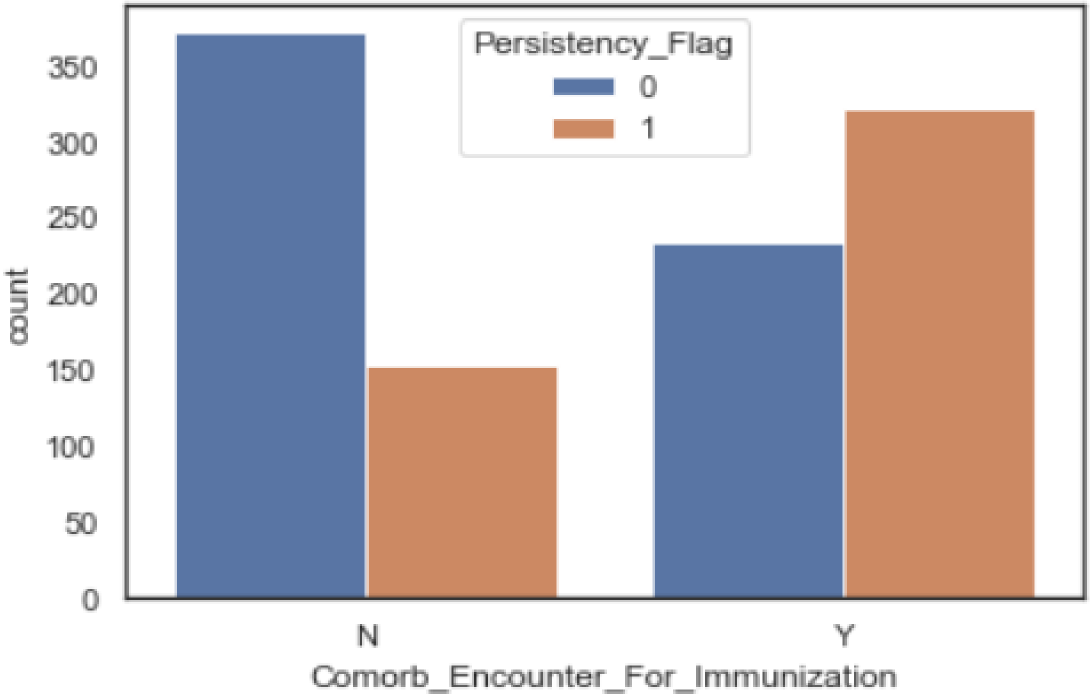
Hypotethis & Data Visualization

Number of DEXA scans by each region

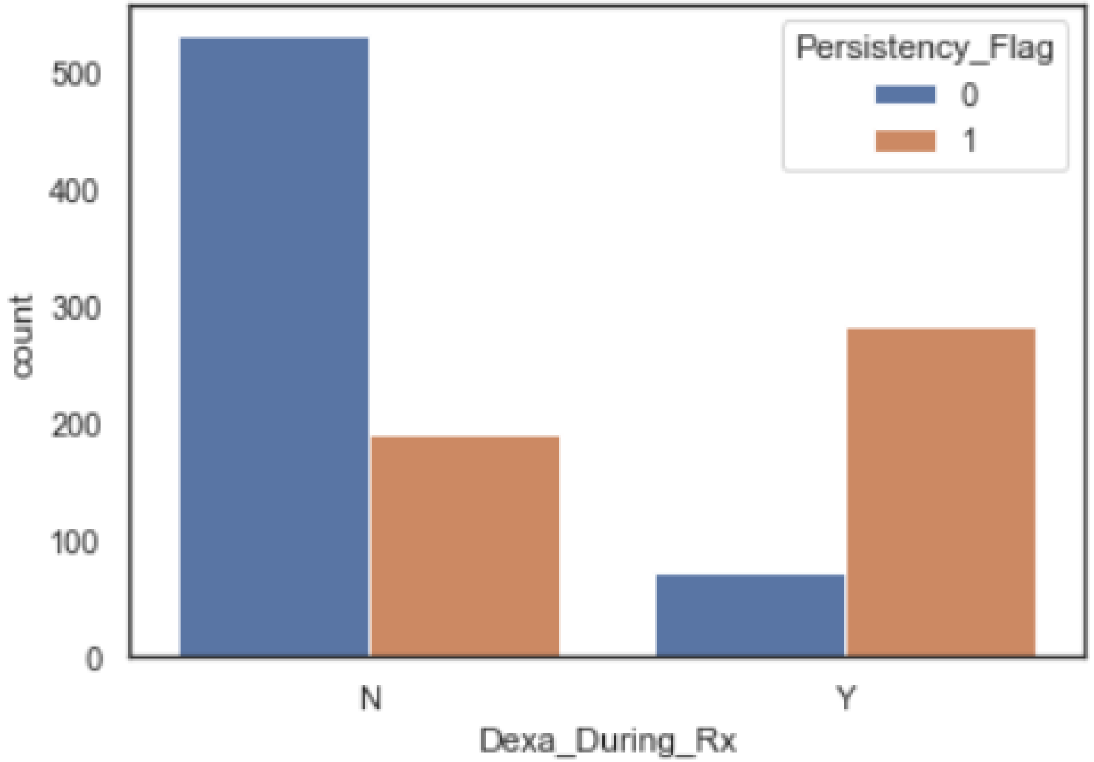


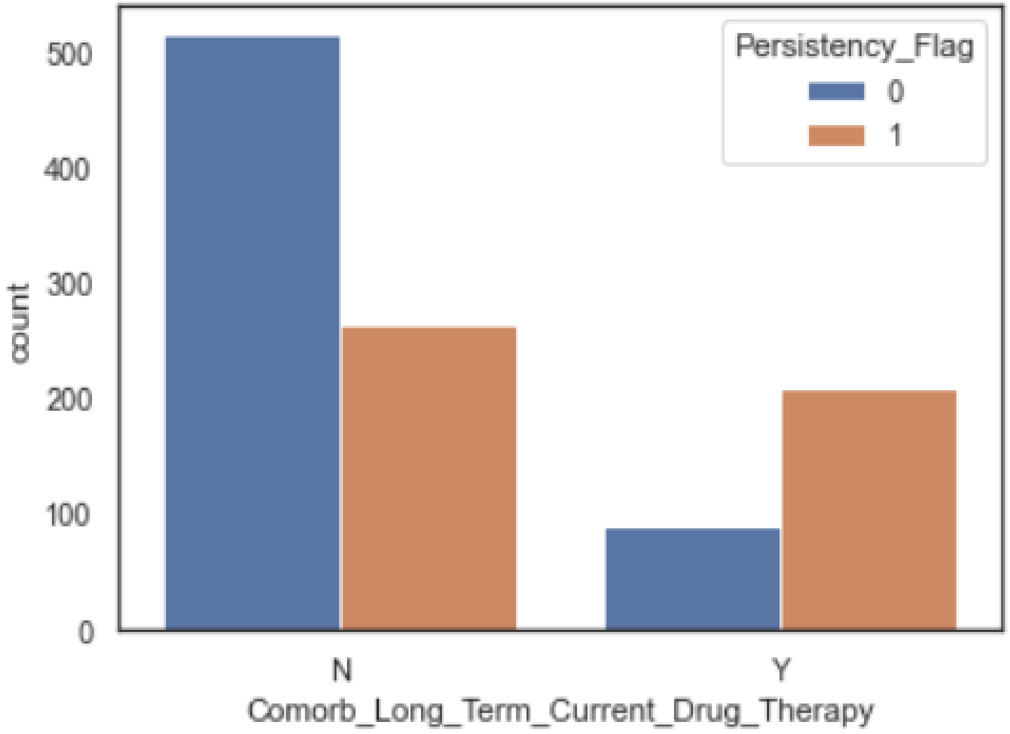
Hypotethis & Data Visualization





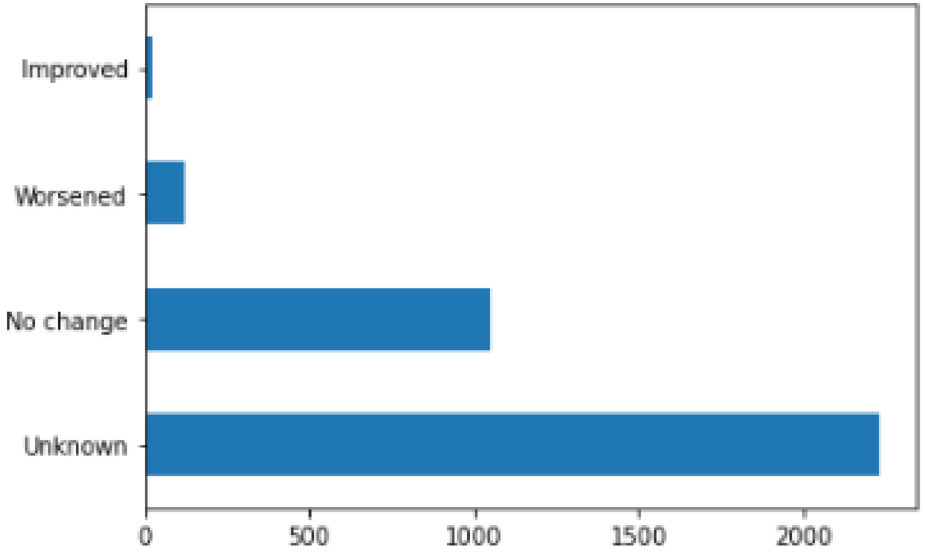
Hypotethis & Data Visualization



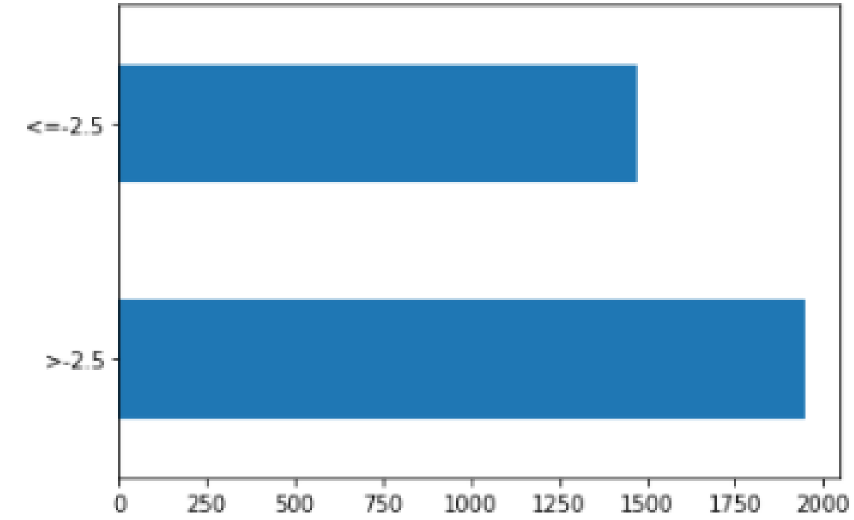


There are similar result counts on features: ‘Dexa_During_Rx’, ‘Comorb_Long_Term_Current_Drug_Therapy’ and ‘Comorb_Encounter_For_Immunization’ has more persistency on ‘yes’ values.

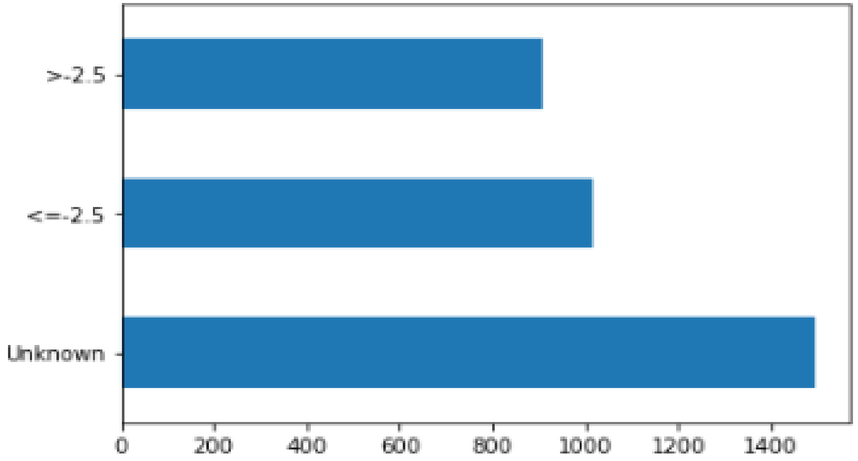
Clinical Factors and T-Scores



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

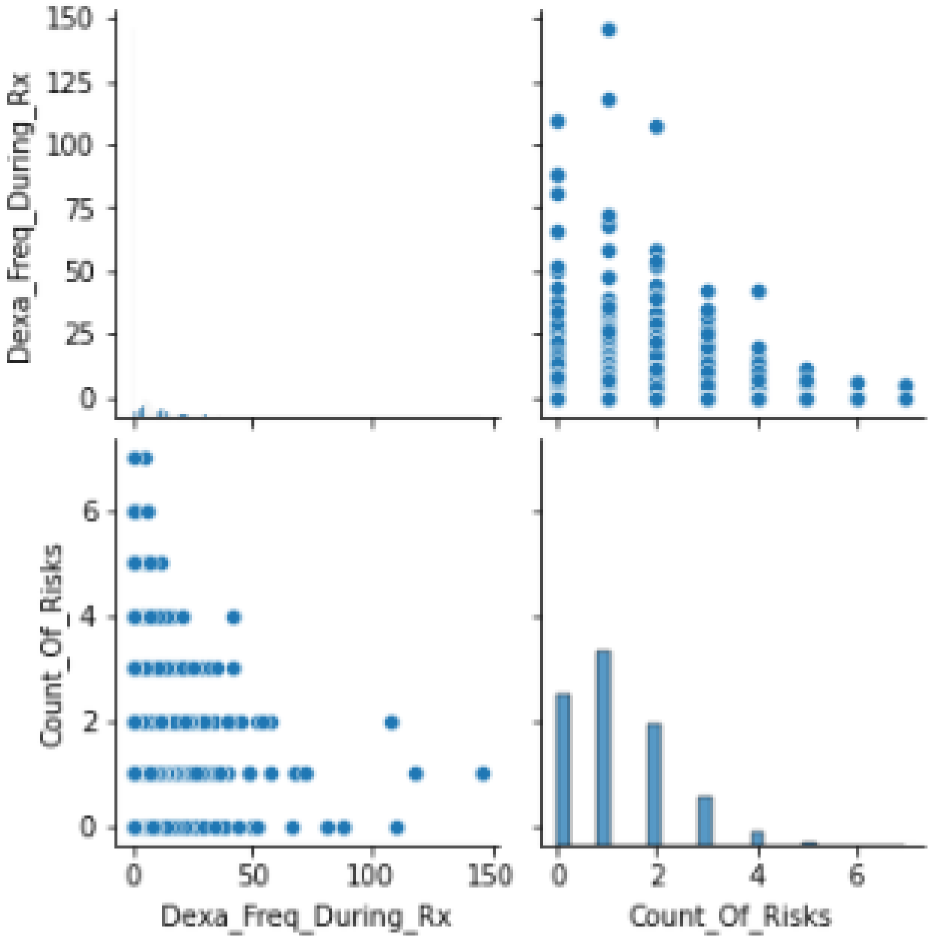


We have compared the "T-scores". The following picture shows the prior to NTM:



T-scores during the Rx:

Numerical Values



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

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
```

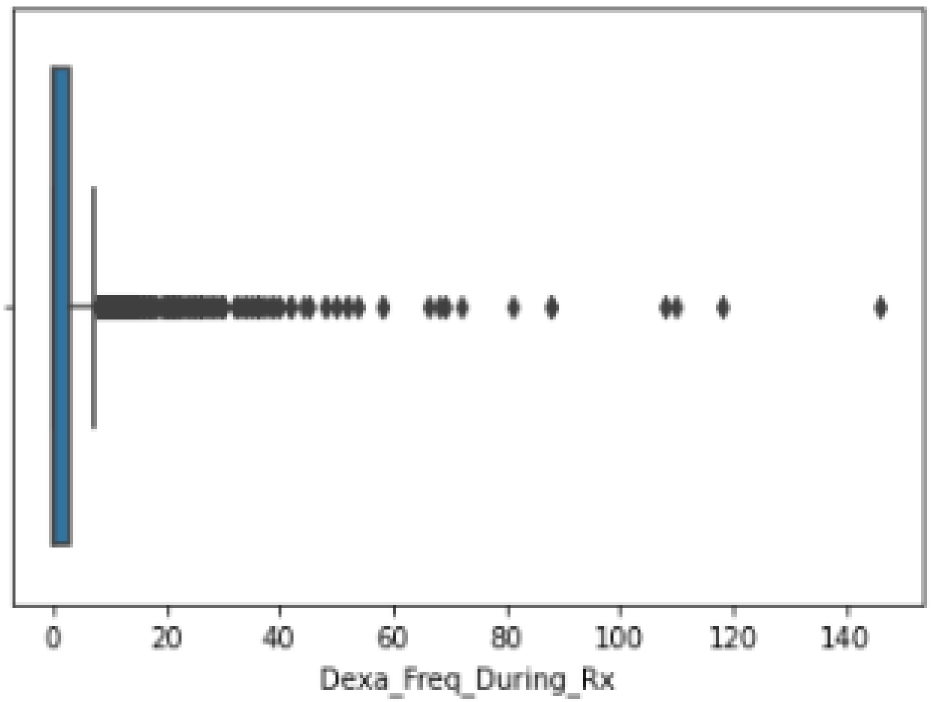
```
In [20]: df["Risk_Segment_During_Rx"].value_counts()

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

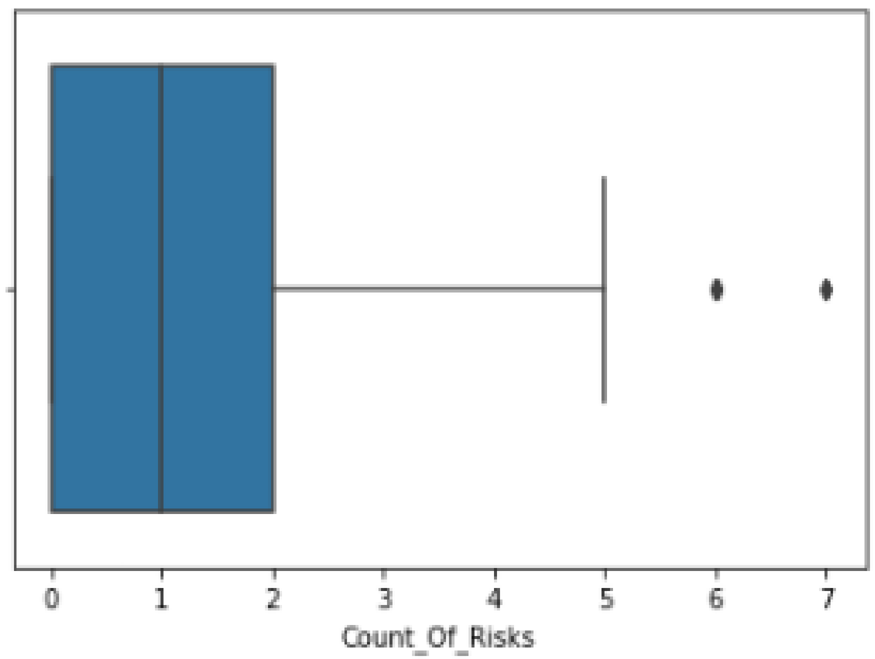
On the other hand, we found a lot of the "Unknown" values, we considered them as null values and decided to remove them because they can affect the results of our ML models.

Outliers

We have 460 outliers in "Dexa_Freq_During_Rx" variable.



We have 8 outliers in "Count_Of_Risks" variable.

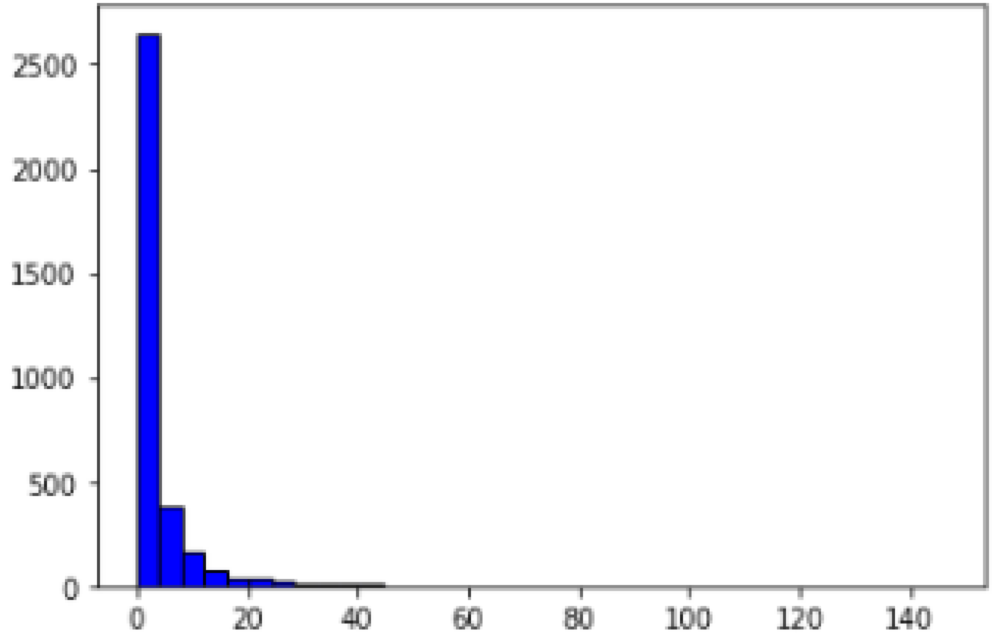


Skewed Data

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:



Recommendations

We have applied Logistic Regression, KNN, Random Forest Model, Neural Network, Gradient Boosting Model, Support Vector Machines and Classification Trees Models.

As seen in the above, we have compared their accuracy scores and we obtained the following:

Gradient Boosting Model is the best fit model to our dataset with accuracy score 0.80.

We can also apply Random Forest Model with accuracy score 0.79 and also we can use Logistic Regression model with accuracy score 0.79 and cross validated score with 10 splits 0.78.

Data Glacier

Your Deep Learning Partner

Your Deep Learning Partner

Thank You