Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Project : Healthcare - Persistency of a drug

Submitted by : Amima Shifa

Agenda

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Problem Statement

One of the challenges for Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. 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, to automate this process of identification.

Objective:

To gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

Target Variable:

Persistency_Flag

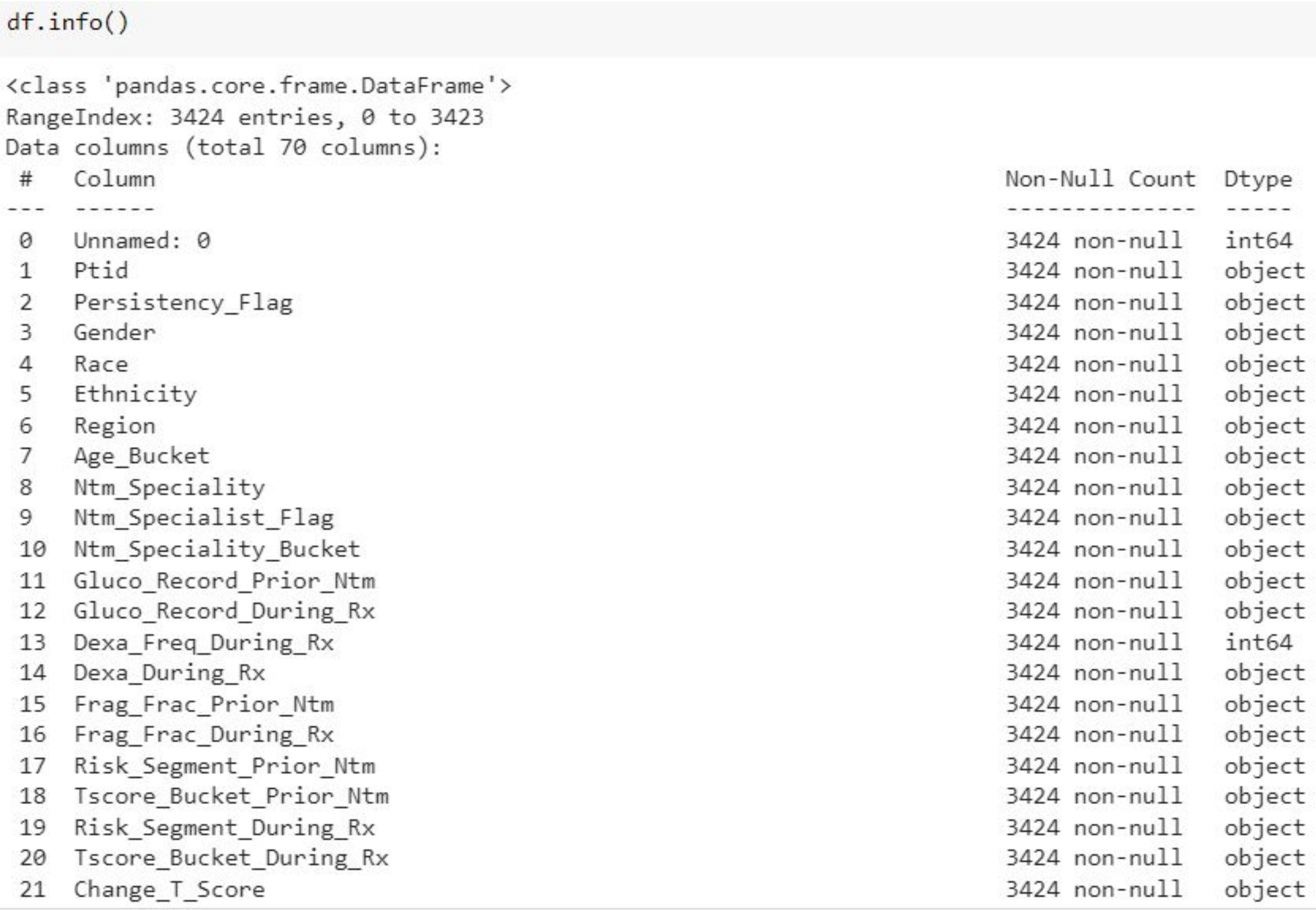
Approach

The approach is divided into the following sections:

- Problem Understanding
- Data Understanding
- Data Cleaning and Transformation
- Data Evaluation
- Data Analysis
- Recommendation

Data Summary

- 70 Features
- 3424 Observations
- Size of data : 898 kb



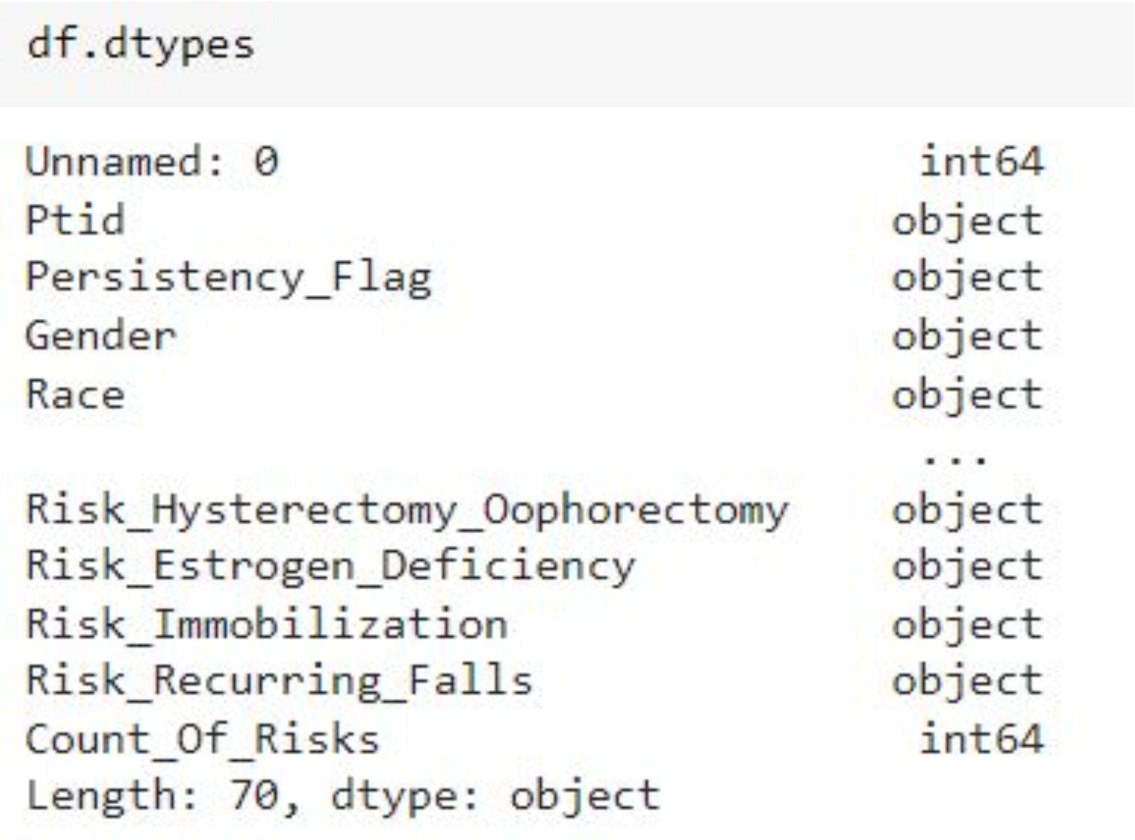
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3424 entries, 0 to 3423
Data columns (total 70 columns):
 #   Column                                                     Non-Null Count  Dtype
---  ------                                                     --------------  -----
 0   Unnamed: 0                                                 3424 non-null   int64
 1   Ptid                                                       3424 non-null   object
 2   Persistency_Flag                                           3424 non-null   object
 3   Gender                                                     3424 non-null   object
 4   Race                                                       3424 non-null   object
 5   Ethnicity                                                  3424 non-null   object
 6   Region                                                     3424 non-null   object
 7   Age_Bucket                                                 3424 non-null   object
 8   Ntm_Speciality                                             3424 non-null   object
 9   Ntm_Specialist_Flag                                        3424 non-null   object
 10  Ntm_Speciality_Bucket                                      3424 non-null   object
 11  Gluco_Record_Prior_Ntm                                     3424 non-null   object
 12  Gluco_Record_During_Rx                                     3424 non-null   object
 13  Dexa_Freq_During_Rx                                        3424 non-null   int64
 14  Dexa_During_Rx                                             3424 non-null   object
 15  Frag_Frac_Prior_Ntm                                        3424 non-null   object
 16  Frag_Frac_During_Rx                                        3424 non-null   object
 17  Risk_Segment_Prior_Ntm                                     3424 non-null   object
 18  Tscore_Bucket_Prior_Ntm                                    3424 non-null   object
 19  Risk_Segment_During_Rx                                     3424 non-null   object
 20  Tscore_Bucket_During_Rx                                    3424 non-null   object
 21  Change_T_Score                                             3424 non-null   object
```

Data Glacier
Your Deep Learning Partner

EDA

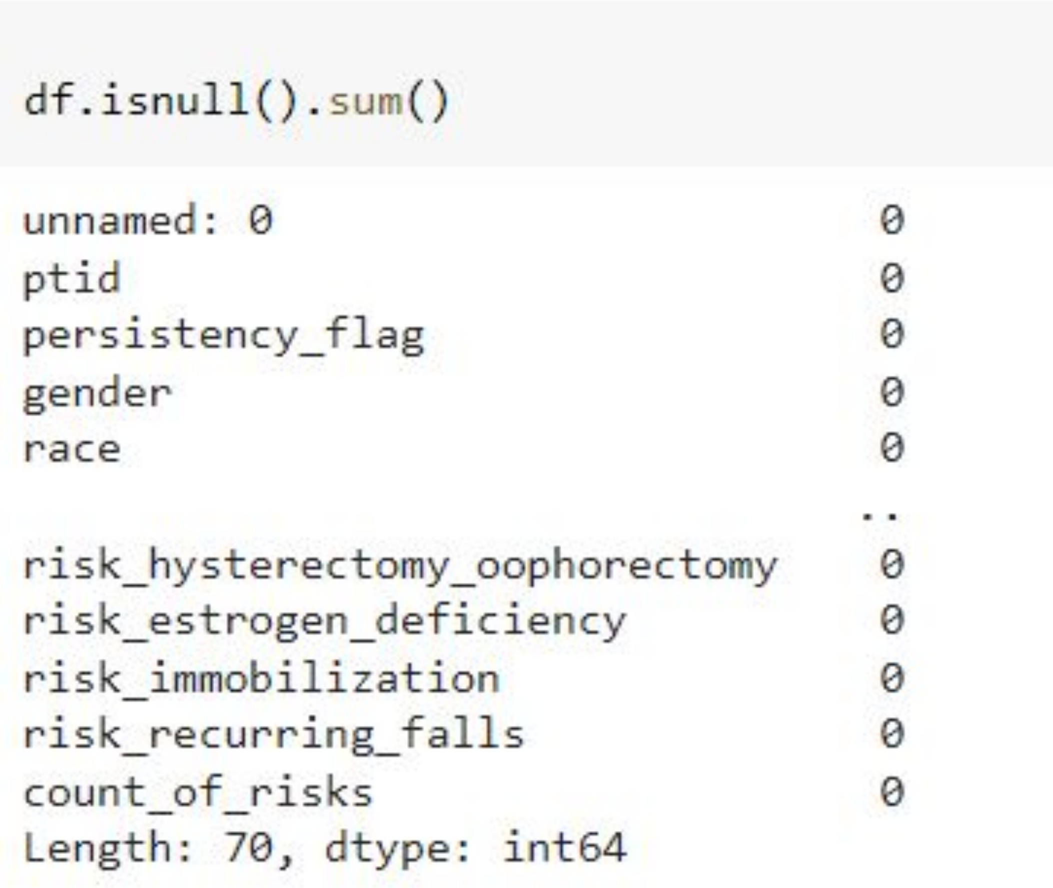
Data types



```
df.dtypes
```

```
Unnamed: 0                            int64
Ptid                                 object
Persistency_Flag                     object
Gender                               object
Race                                 object
                                      ...
Risk_Hysterectomy_Oophorectomy       object
Risk_Estrogen_Deficiency             object
Risk_Immobilization                  object
Risk_Recurring_Falls                 object
Count_Of_Risks                        int64
Length: 70, dtype: object
```

Missing Values



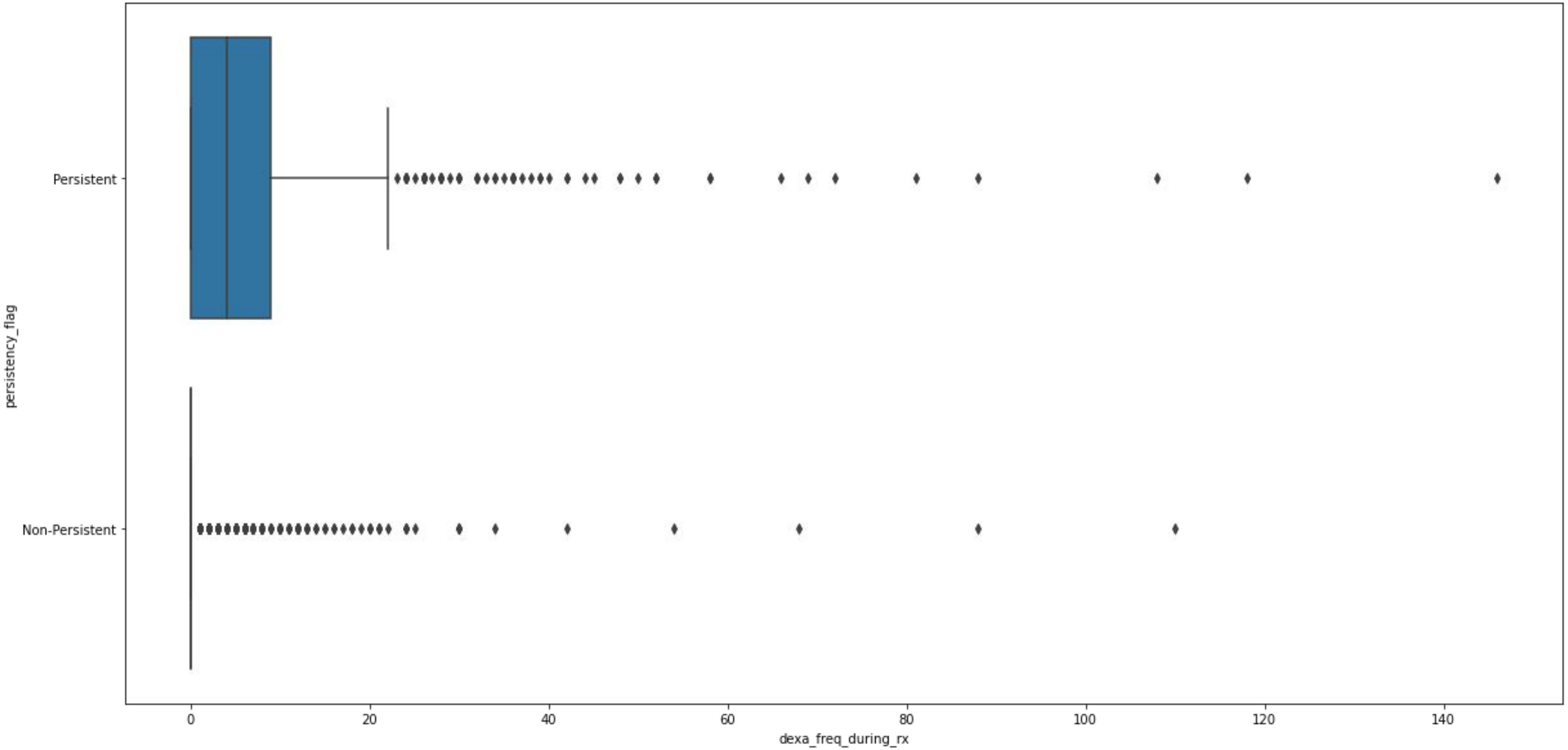
```
df.isnull().sum()
```

```
unnamed: 0                           0
ptid                                 0
persistency_flag                     0
gender                               0
race                                 0
                                    ..
risk_hysterectomy_oophorectomy       0
risk_estrogen_deficiency             0
risk_immobilization                  0
risk_recurring_falls                 0
count_of_risks                       0
Length: 70, dtype: int64
```

There are no missing values present in the dataset.

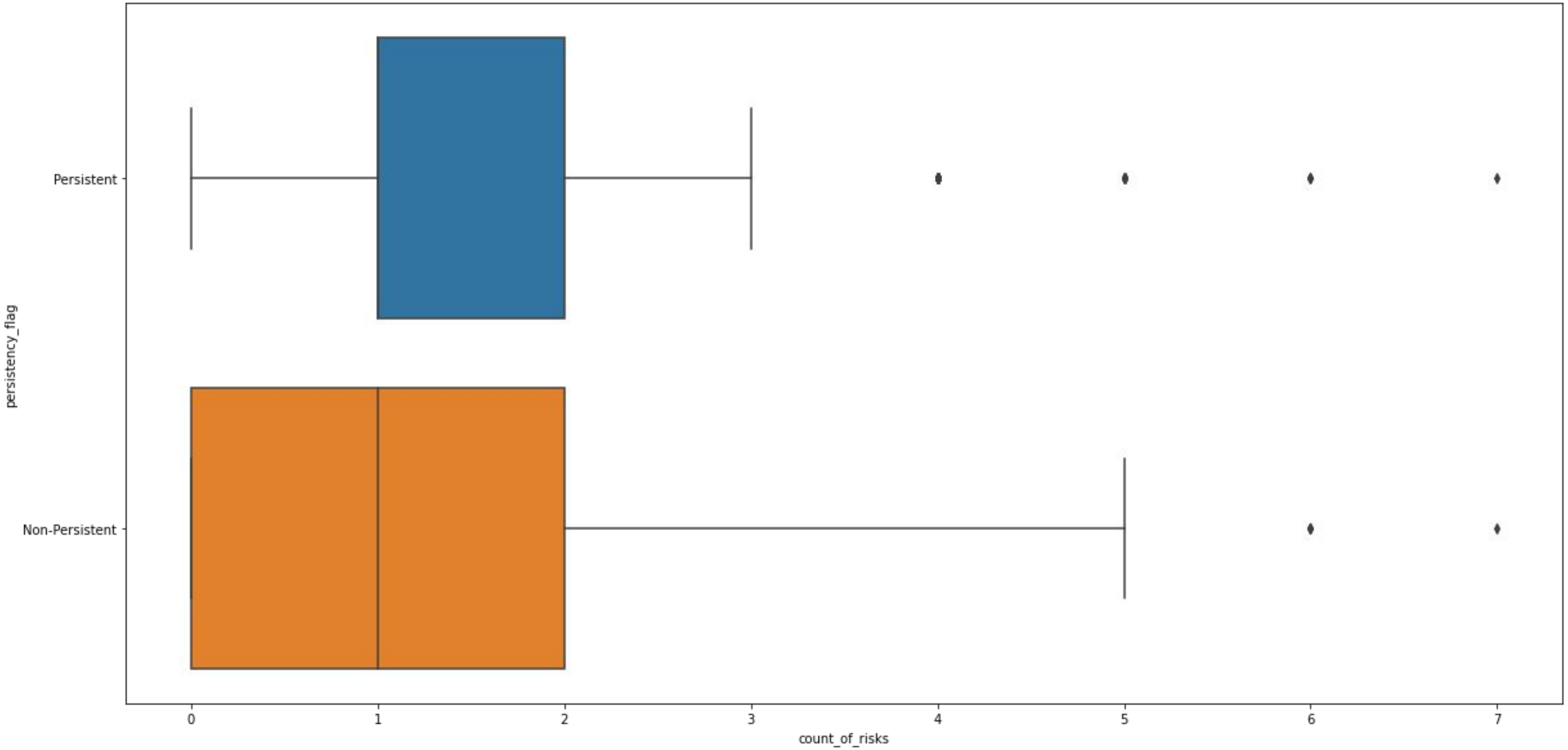
Data Glacier
Your Deep Learning Partner

Outlier Analysis



Outliers are present in Dexa Frequency during RX .

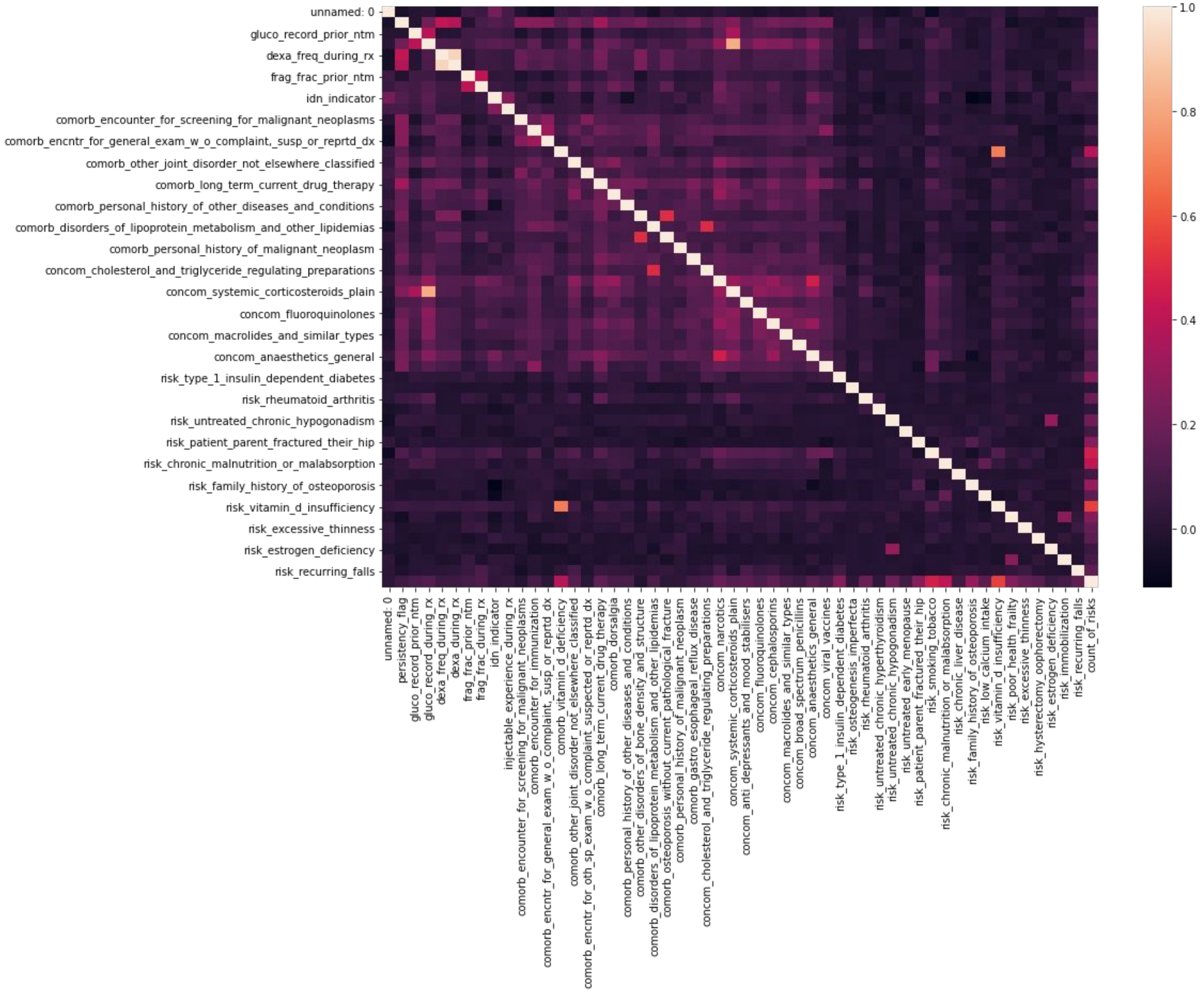
Outlier Analysis



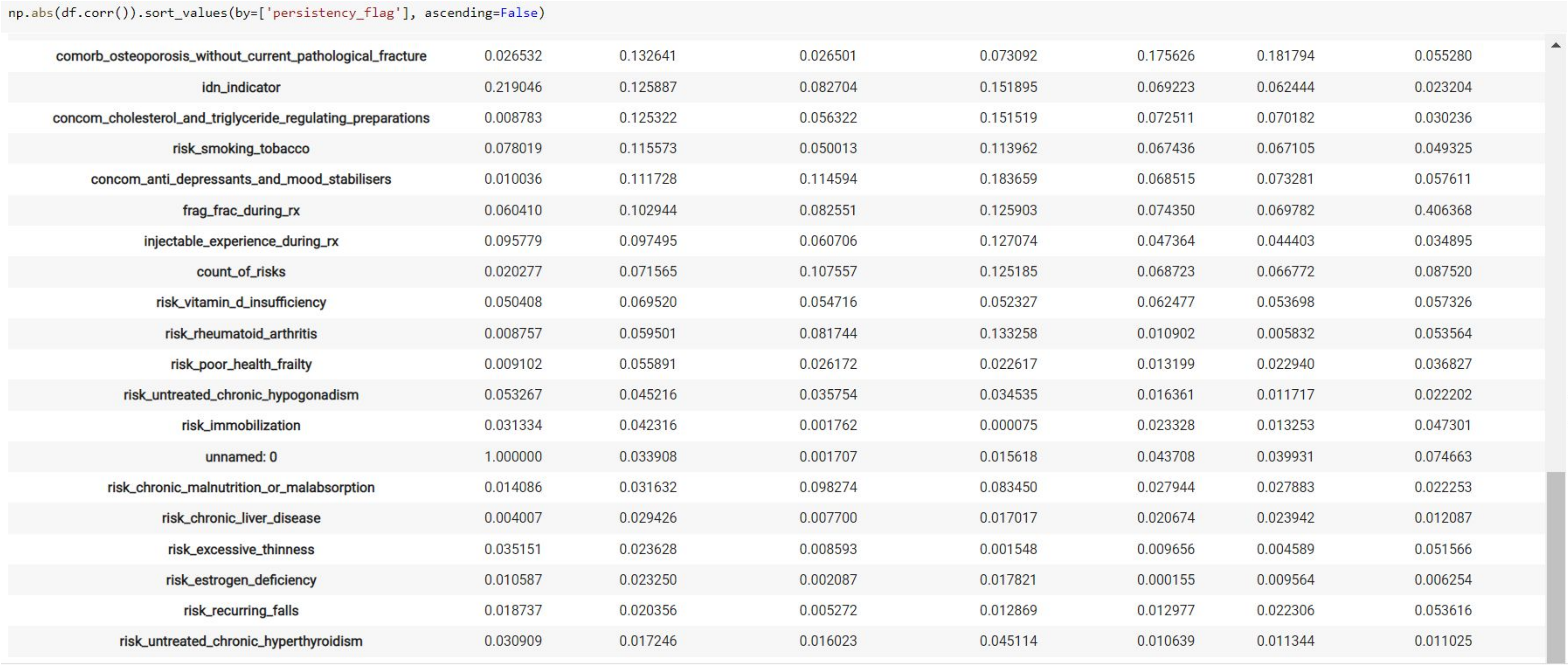
Outliers are present in Count of Risks.

Data Glacier
Your Deep Learning Partner

Correlation Analysis
(After Transformation)



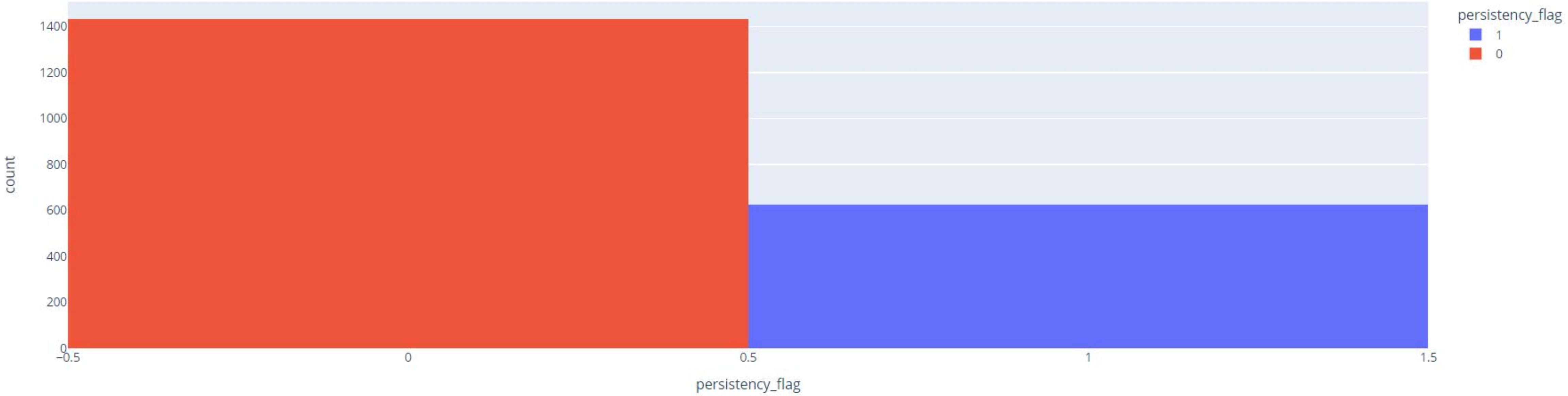
Correlation Analysis
(After Transformation)



```
np.abs(df.corr()).sort_values(by=['persistency_flag'], ascending=False)
```

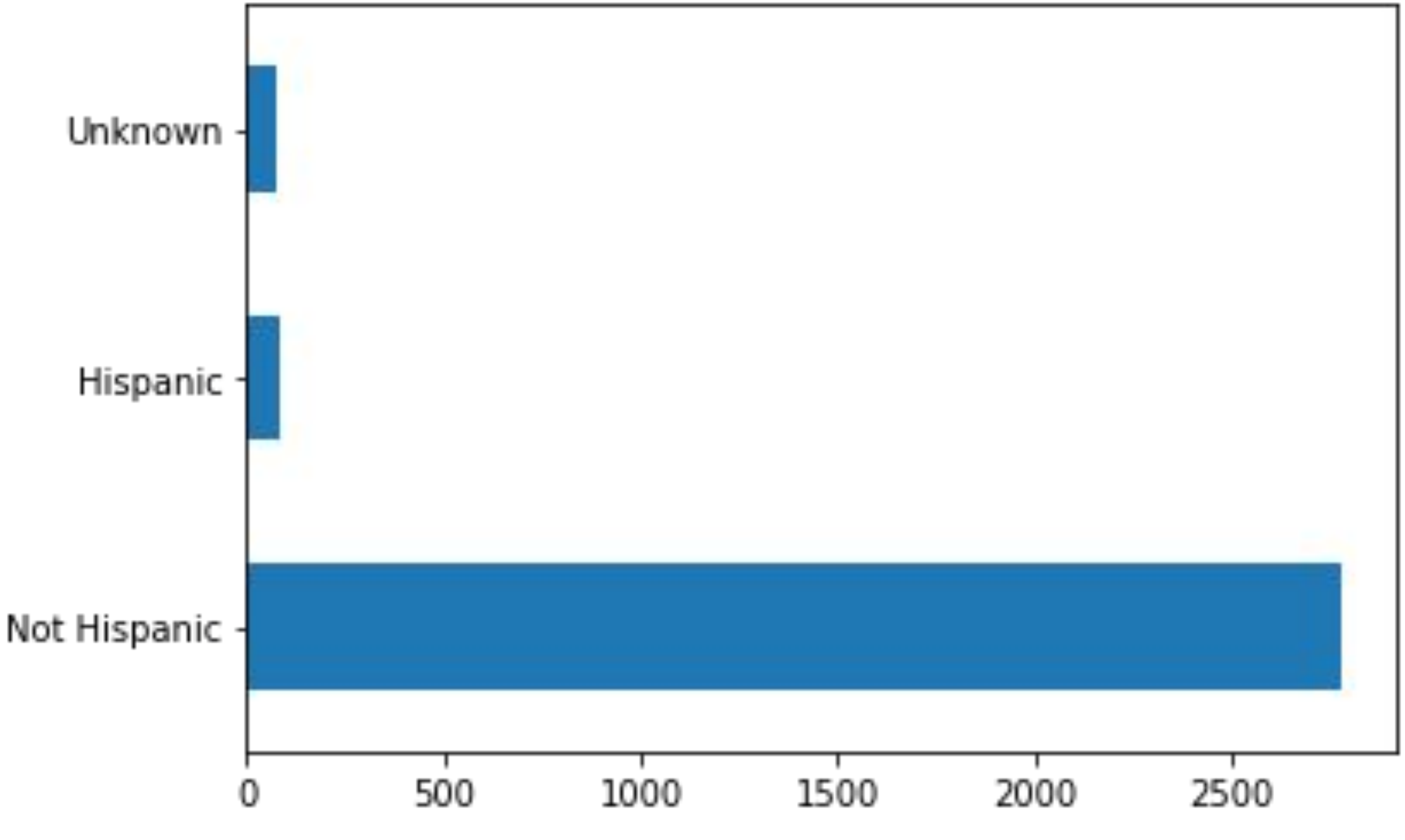
comorb_osteoporosis_without_current_pathological_fracture	0.026532	0.132641	0.026501	0.073092	0.175626	0.181794	0.055280
idn_indicator	0.219046	0.125887	0.082704	0.151895	0.069223	0.062444	0.023204
concom_cholesterol_and_triglyceride_regulating_preparations	0.008783	0.125322	0.056322	0.151519	0.072511	0.070182	0.030236
risk_smoking_tobacco	0.078019	0.115573	0.050013	0.113962	0.067436	0.067105	0.049325
concom_anti_depressants_and_mood_stabilisers	0.010036	0.111728	0.114594	0.183659	0.068515	0.073281	0.057611
frag_frac_during_rx	0.060410	0.102944	0.082551	0.125903	0.074350	0.069782	0.406368
injectable_experience_during_rx	0.095779	0.097495	0.060706	0.127074	0.047364	0.044403	0.034895
count_of_risks	0.020277	0.071565	0.107557	0.125185	0.068723	0.066772	0.087520
risk_vitamin_d_insufficiency	0.050408	0.069520	0.054716	0.052327	0.062477	0.053698	0.057326
risk_rheumatoid_arthritis	0.008757	0.059501	0.081744	0.133258	0.010902	0.005832	0.053564
risk_poor_health_frailty	0.009102	0.055891	0.026172	0.022617	0.013199	0.022940	0.036827
risk_untreated_chronic_hypogonadism	0.053267	0.045216	0.035754	0.034535	0.016361	0.011717	0.022202
risk_immobilization	0.031334	0.042316	0.001762	0.000075	0.023328	0.013253	0.047301
unnamed: 0	1.000000	0.033908	0.001707	0.015618	0.043708	0.039931	0.074663
risk_chronic_malnutrition_or_malabsorption	0.014086	0.031632	0.098274	0.083450	0.027944	0.027883	0.022253
risk_chronic_liver_disease	0.004007	0.029426	0.007700	0.017017	0.020674	0.023942	0.012087
risk_excessive_thinness	0.035151	0.023628	0.008593	0.001548	0.009656	0.004589	0.051566
risk_estrogen_deficiency	0.010587	0.023250	0.002087	0.017821	0.000155	0.009564	0.006254
risk_recurring_falls	0.018737	0.020356	0.005272	0.012869	0.012977	0.022306	0.053616
risk_untreated_chronic_hyperthyroidism	0.030909	0.017246	0.016023	0.045114	0.010639	0.011344	0.011025

Persistence Flag



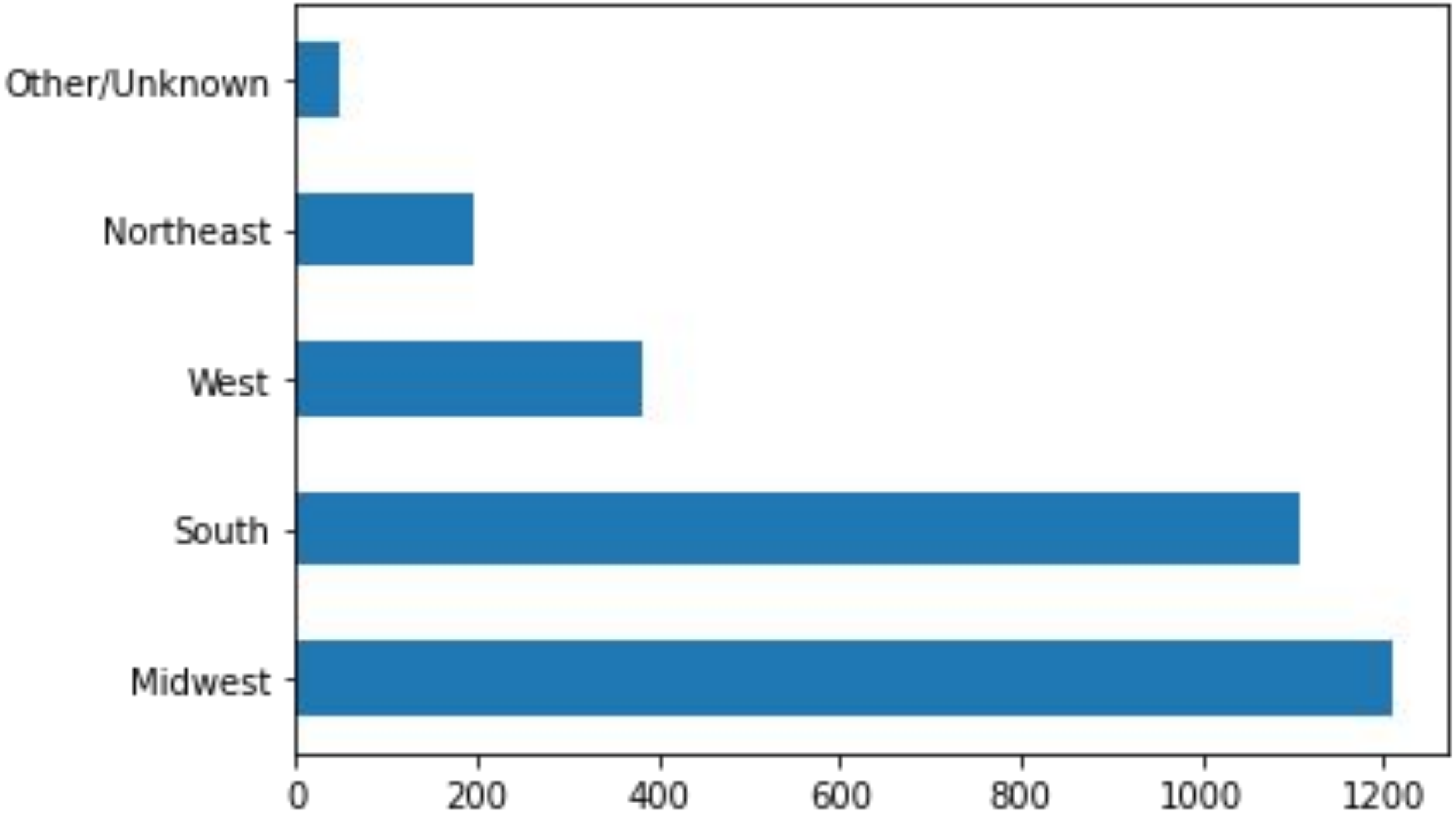
Less drugs are persistent than non-persistent.

Ethnicity Distribution



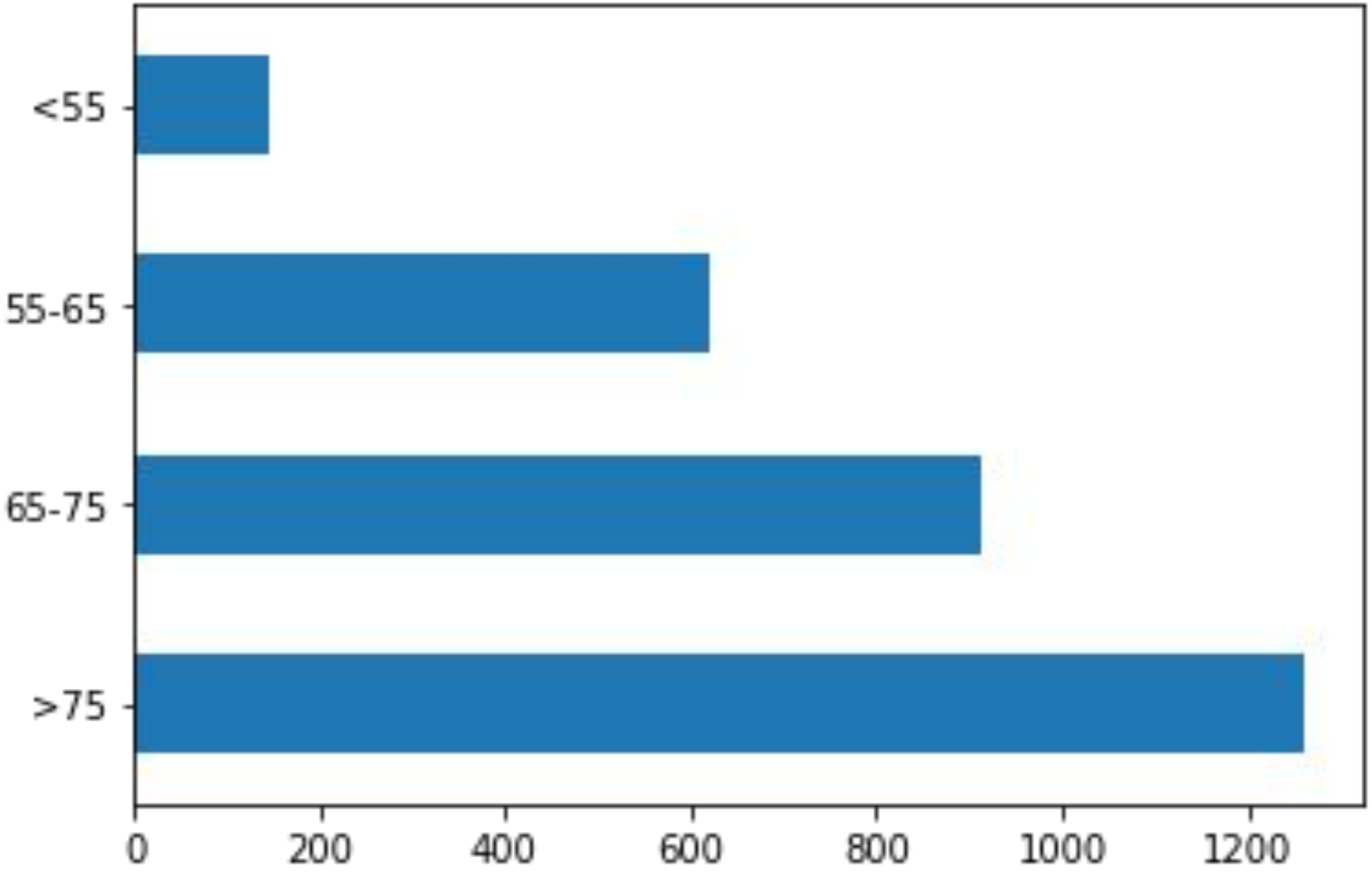
The highest ethnicity is of not hispanic people.

Region wise Distribution



South and Midwest are the dominant regions.

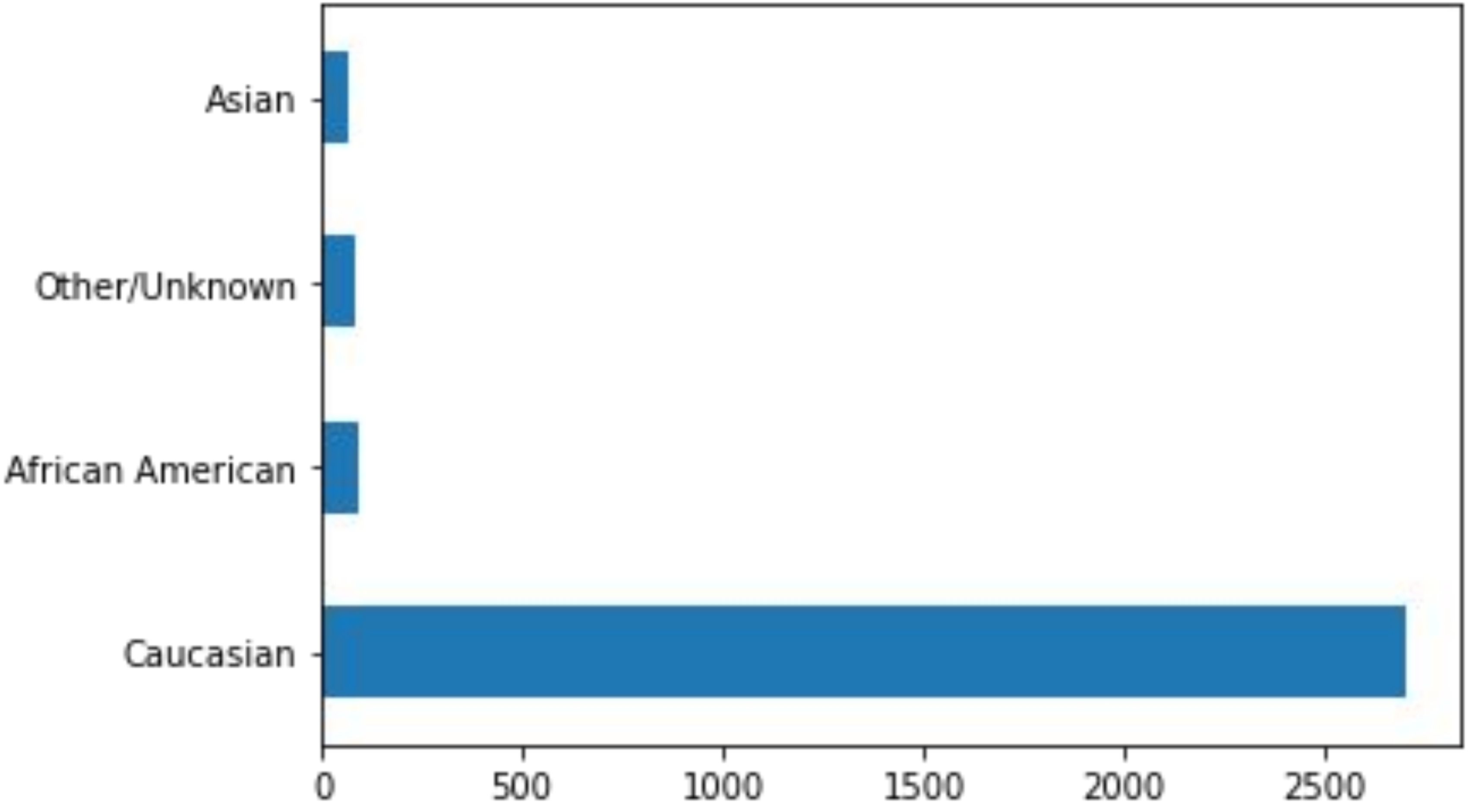
Age analysis



Majority of the patients are above the age of 55 years.

Data Glacier
Your Deep Learning Partner

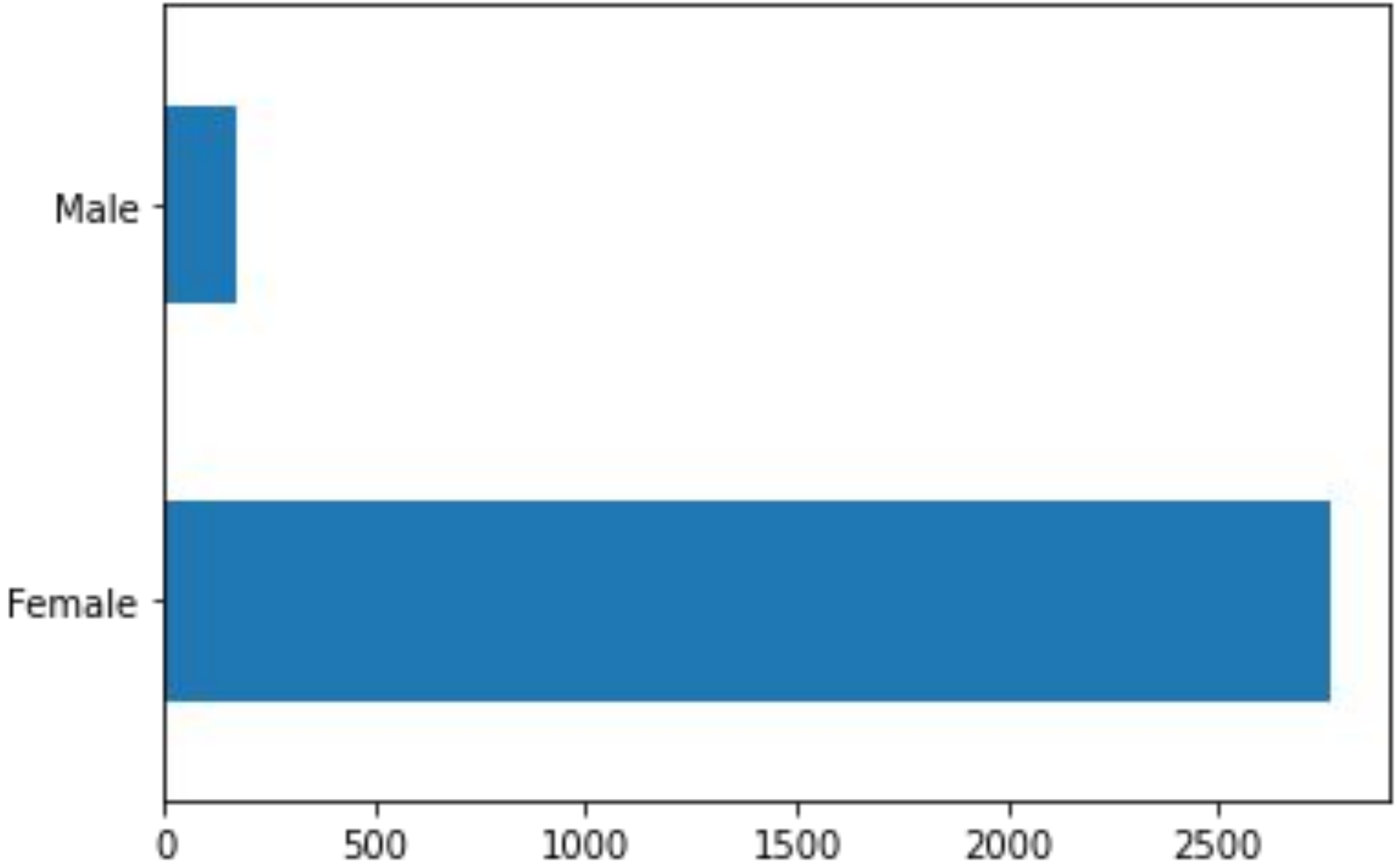
Race Distribution



Caucasian race is the most prominent among other races.

Data Glacier
Your Deep Learning Partner

Gender Analysis

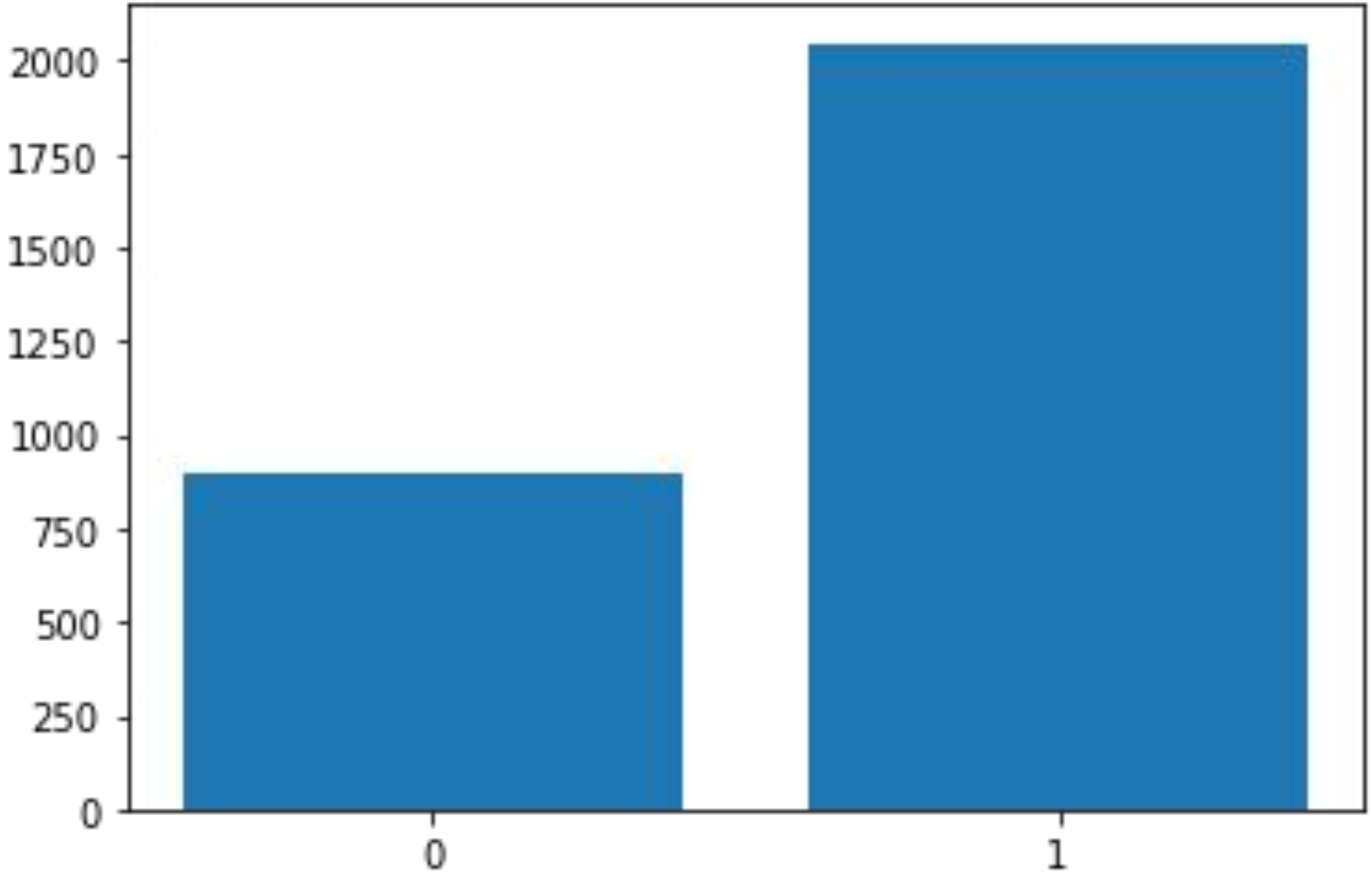


Female patients are considerably more than male patients.

Data Glacier

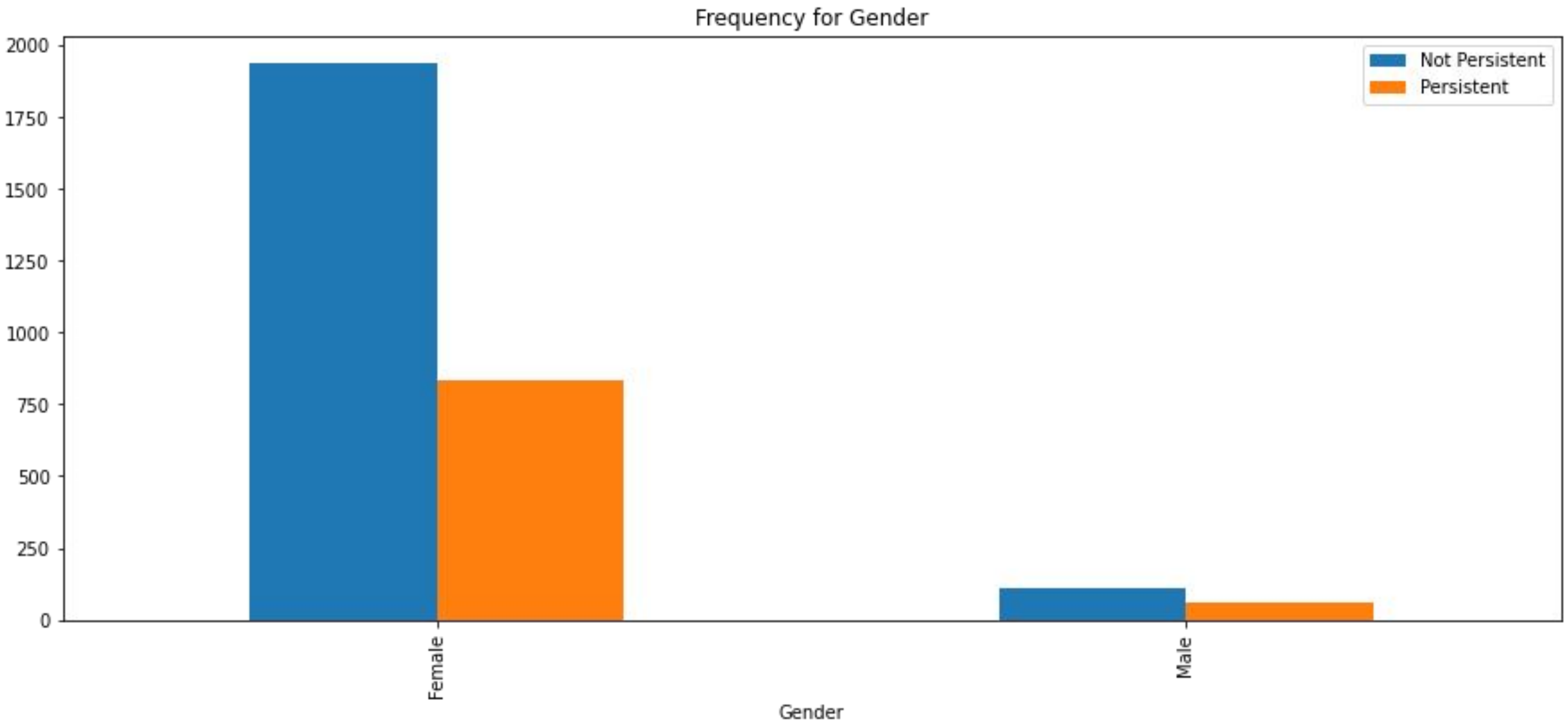
Your Deep Learning Partner

Persistency Flag



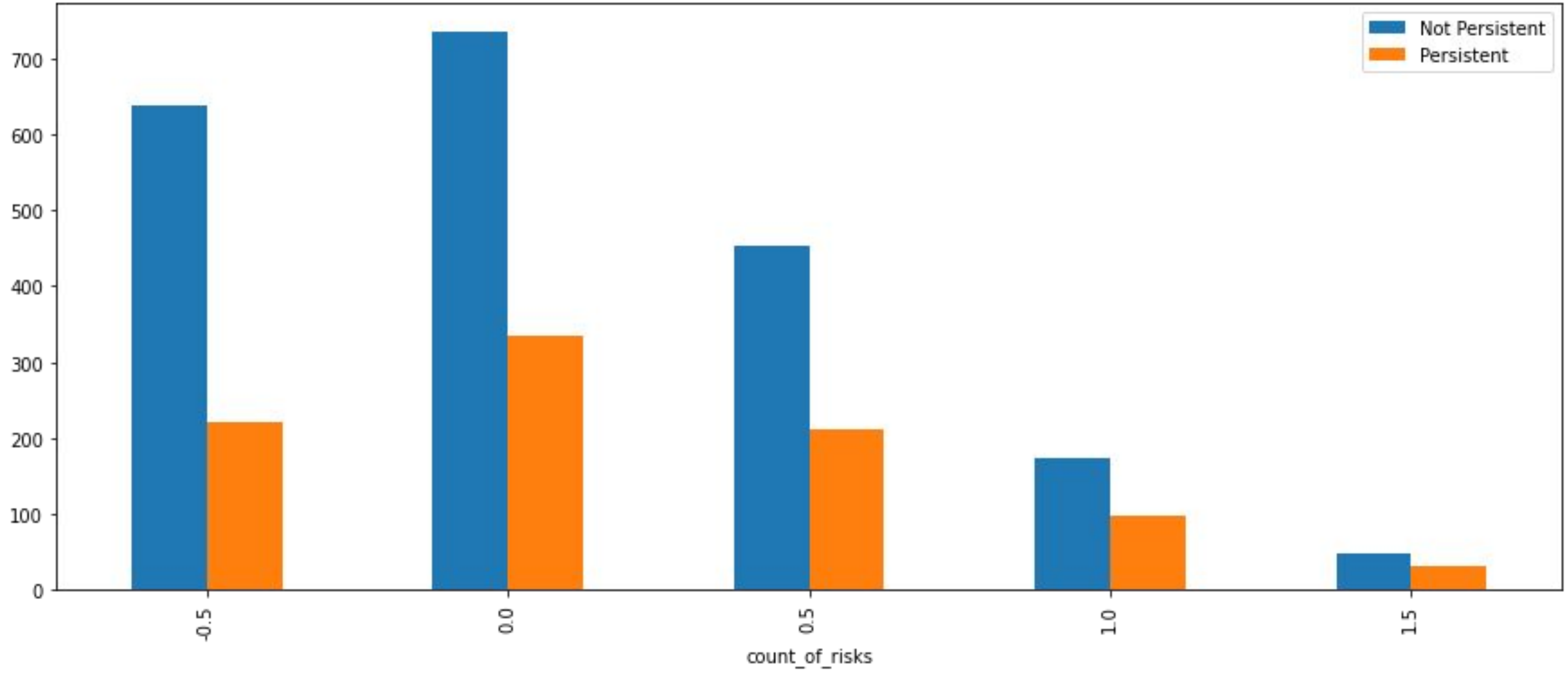
Drugs are more persistent than non-persistent.

Persistency with respect to Gender



Females and males are mostly persistent to the drug.

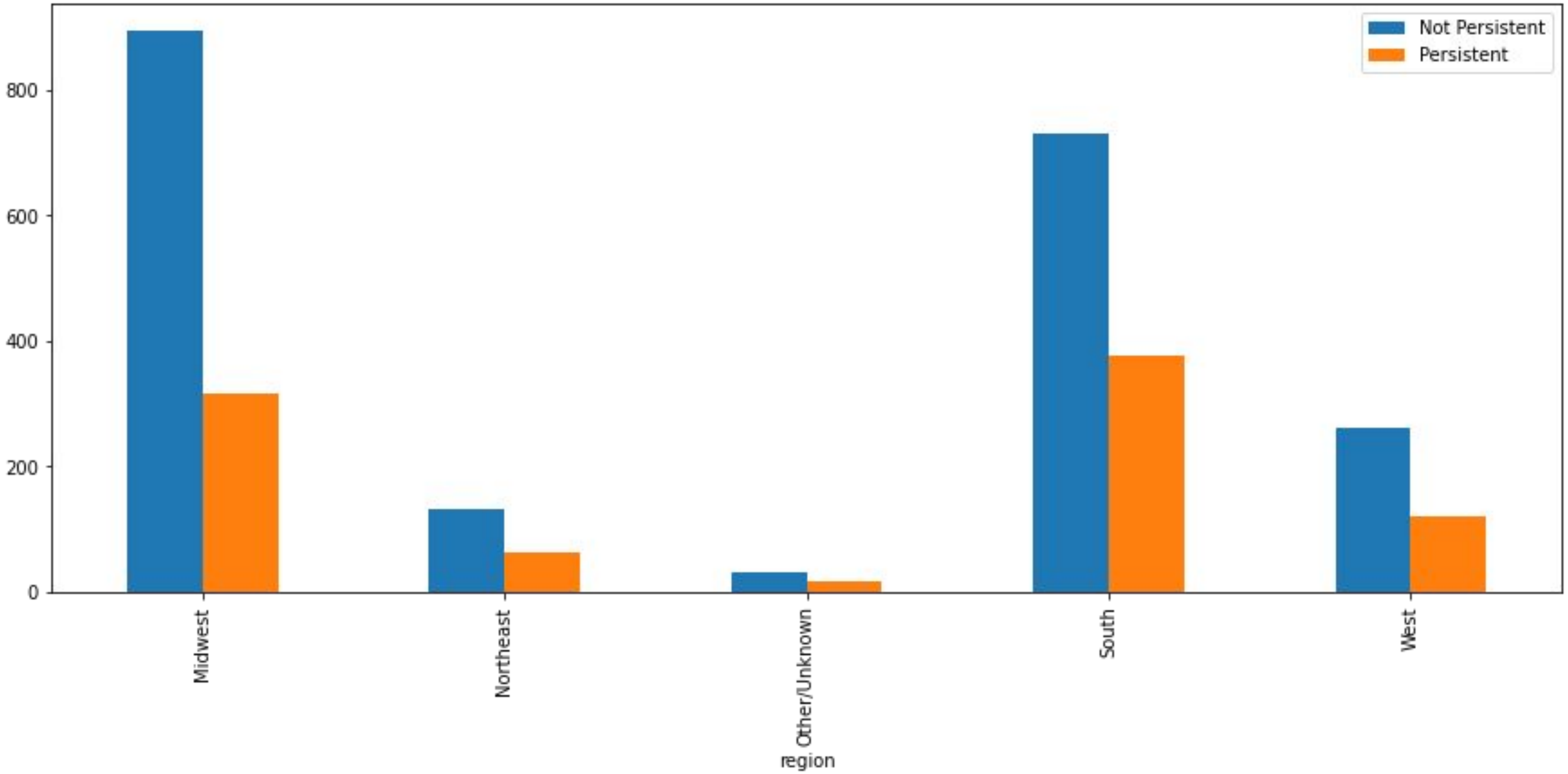
Risk count with respect to persistency



Number of risks with non-persistent drug is larger than with persistent drug.

Data Glacier
Your Deep Learning Partner

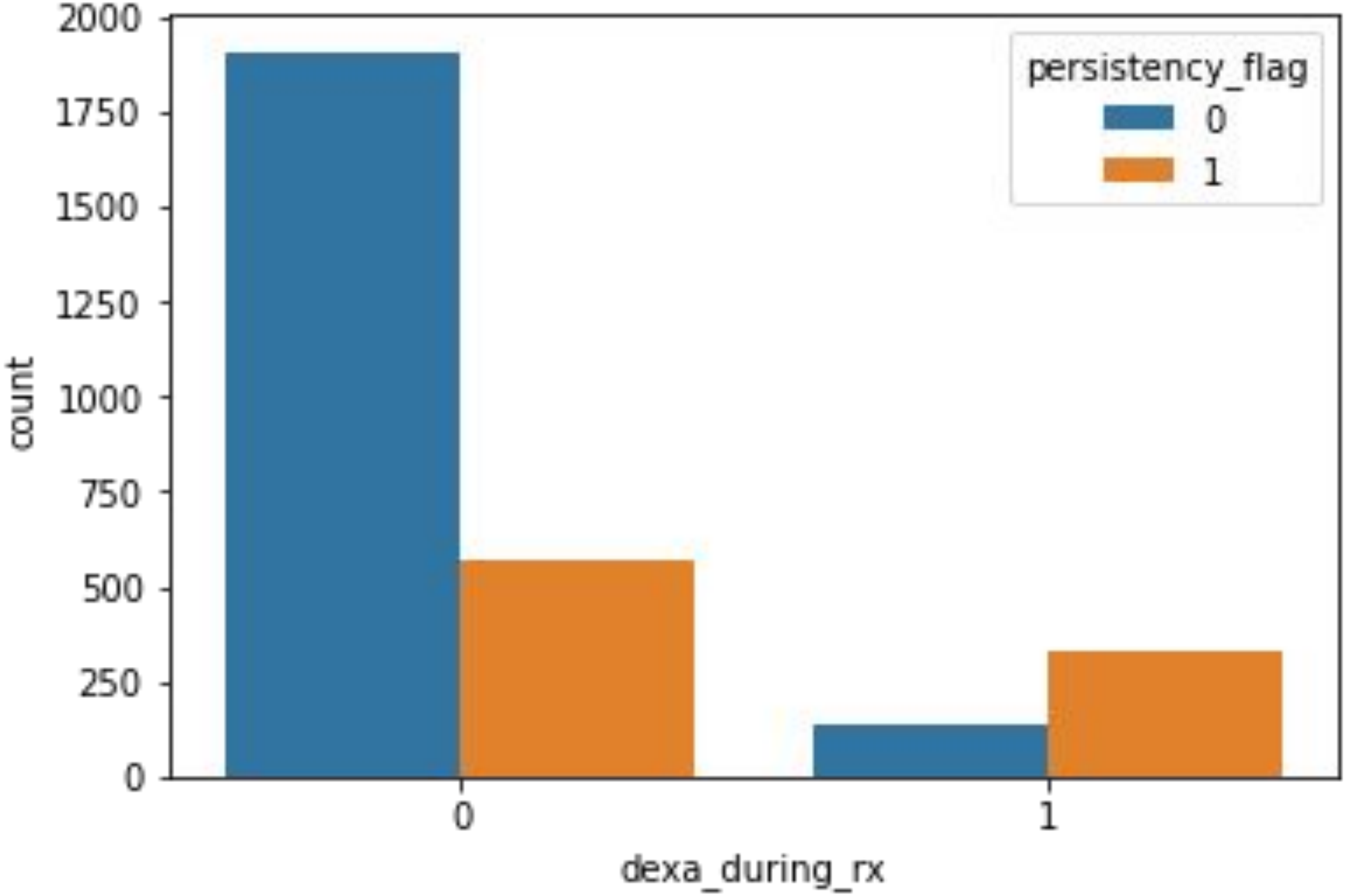
Region wise analysis



For Midwest vast majority of patients show persistence to the drug.

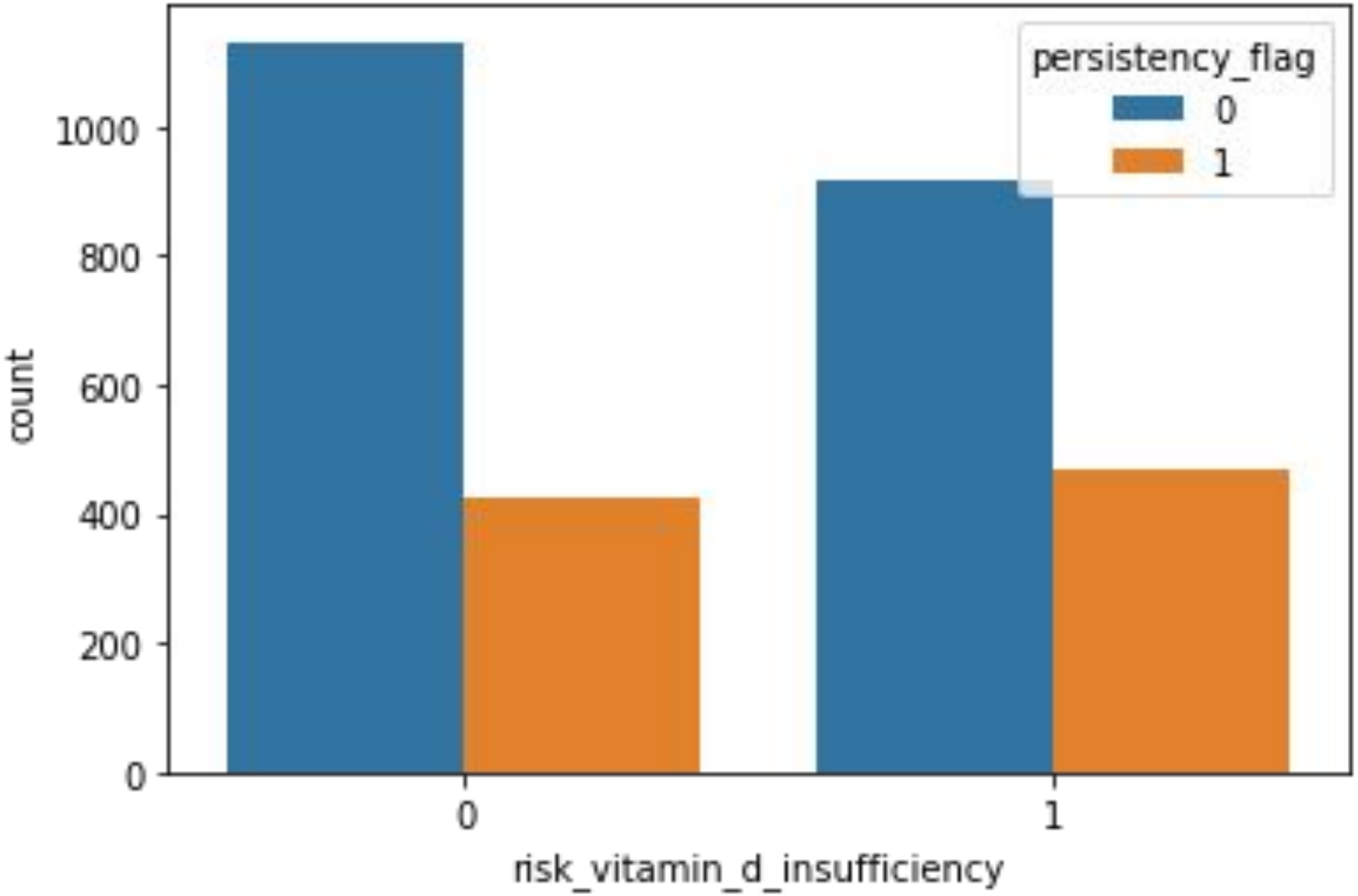
Data Glacier
Your Deep Learning Partner

Patients resistant to drug who had DEXA scan during therapy



Number of patients being persistent to drug with having undergone a DEXA during therapy is high.

Vitamin D Insufficiency Risk



Risk of Vitamin D insufficiency is higher for patients who are non-persistent to the drug.

Conclusion

Recommendation

As it is a classification problem, many different classification models can be used such as Logistic Regression, Random Forest Classifier and Gradient Boosting model.

Thank You

Data Glacier

Your Deep Learning Partner