Exploratory Data Analysis

HEALTH CARE – DRUG PERSISTENCY

DECEMBER 2022

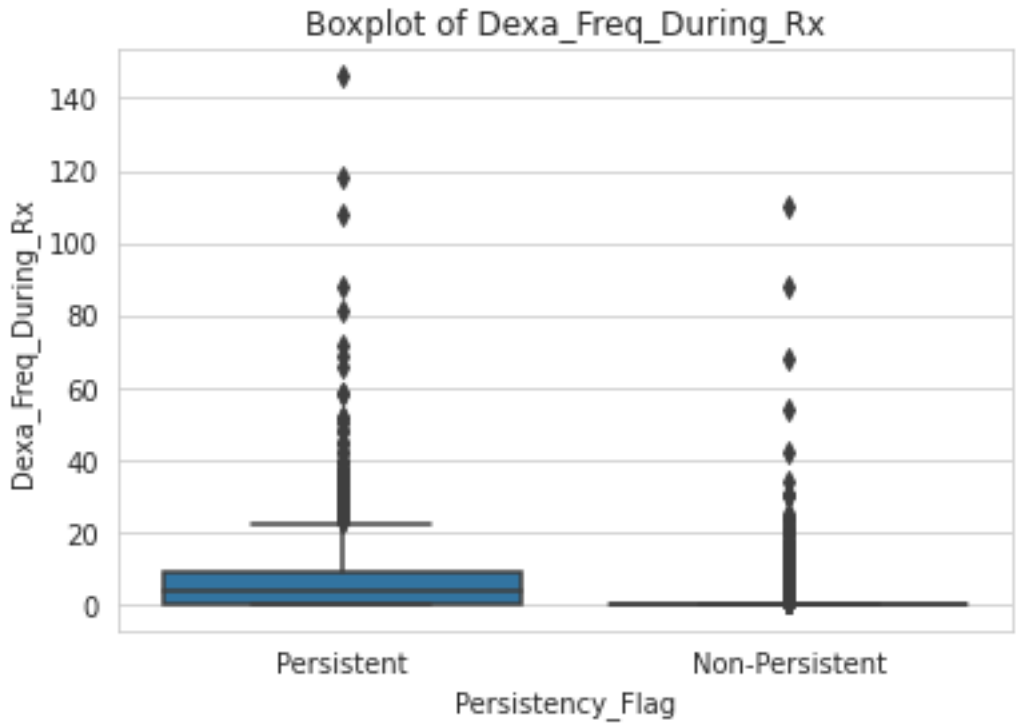
Data Exploration

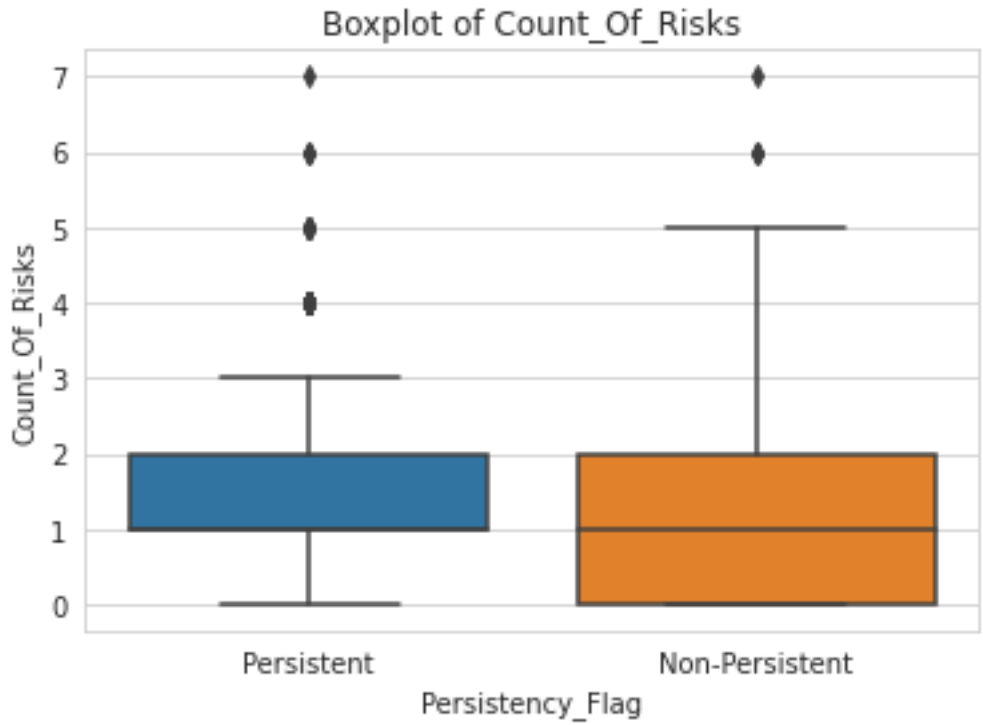
- One file used for the dataset
- 3,424 data points
- 75 features/variables (6 derived)

Analysis of Numerical Features

- Dexa_Freq_During_Rx vs Persistency_Flag
- Count_Of_Risks vs Persistency_Flag

Dexa_Freq_During_Rx and Count_Of_Risks Analysis





- There's a big difference in the distribution of the dexa scan frequency during prescription between persistent patient's and non-persistent patients.
- 50% of persistent patients have count of risks between 1 and 2, while 50% of non-persistent patients have count of risks between 0 and 2.

Numerical Features Correlation

	Persistency_Flag	Dexa_Freq_During_Rx	Count_Of_Risks	log_Dexa	log_Count_Risks
Persistency_Flag	1.000000	0.517337	0.082431	0.517315	0.083250
Dexa_Freq_During_Rx	0.517337	1.000000	0.063414	0.990813	0.064419
Count_Of_Risks	0.082431	0.063414	1.000000	0.067388	0.966552
log_Dexa	0.517315	0.990813	0.067388	1.000000	0.067405
log_Count_Risks	0.083250	0.064419	0.966552	0.067405	1.000000

Analysis of Categorical Features

- MUTUAL INFORMATION (MI) SCORE
- Class separation by categorical features
- Class separation by Dexa_During_Rx
- Class separation by Comorb_Long_Term_Current_Drug_Therapy
- Class separation by Comorb_Encounter_For_Screening_For_Malignant_Neoplasms
- Class separation by Comorb_Encounter_For_Immunization
- Class separation by Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx
- And go on…

MUTUAL INFORMATION (MI) SCORE

Mutual information is the measurement of how much information one can obtain about a random variable given the value of another variable.

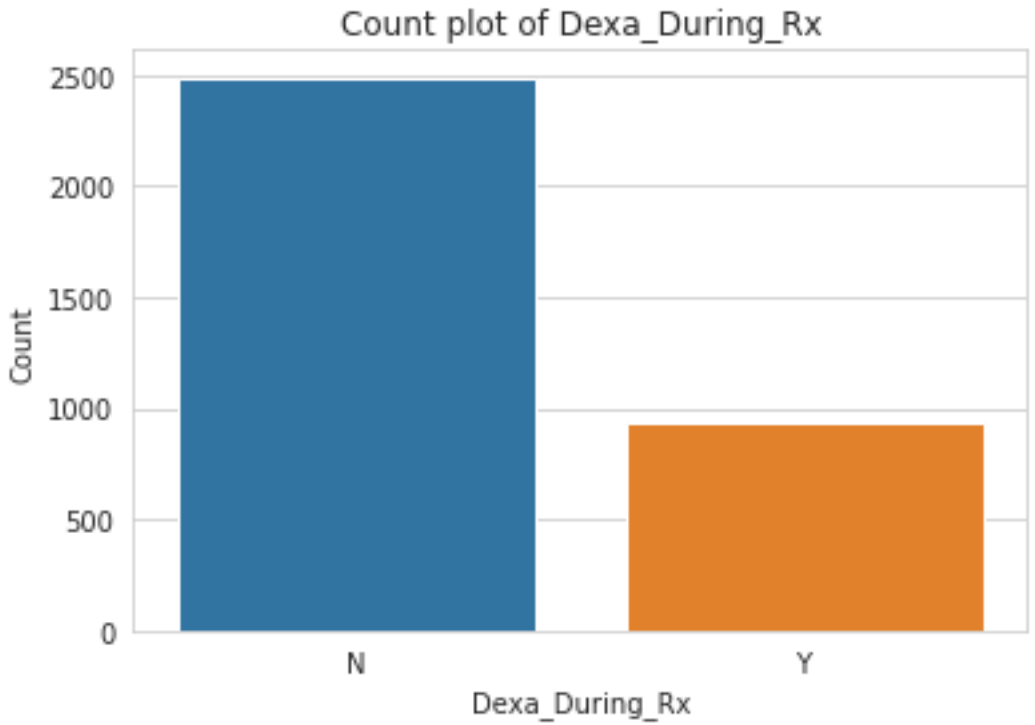
Persistency_Flag	6.623046e-01
Dexa_During_Rx	1.211142e-01
Comorb_Long_Term_Current_Drug_Therapy	6.101576e-02
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	5.251710e-02
Comorb_Encounter_For_Immunization	5.002321e-02
...	...
Gluco_Record_Prior_Ntm	1.659650e-05
Risk_Untreated_Early_Menopause	1.416782e-05
Risk_Family_History_Of_Osteoporosis	6.108238e-06
Risk_Osteogenesis_Imperfecta	3.532791e-06
Frag_Frac_Prior_Ntm	5.052294e-08

66 rows × 1 columns

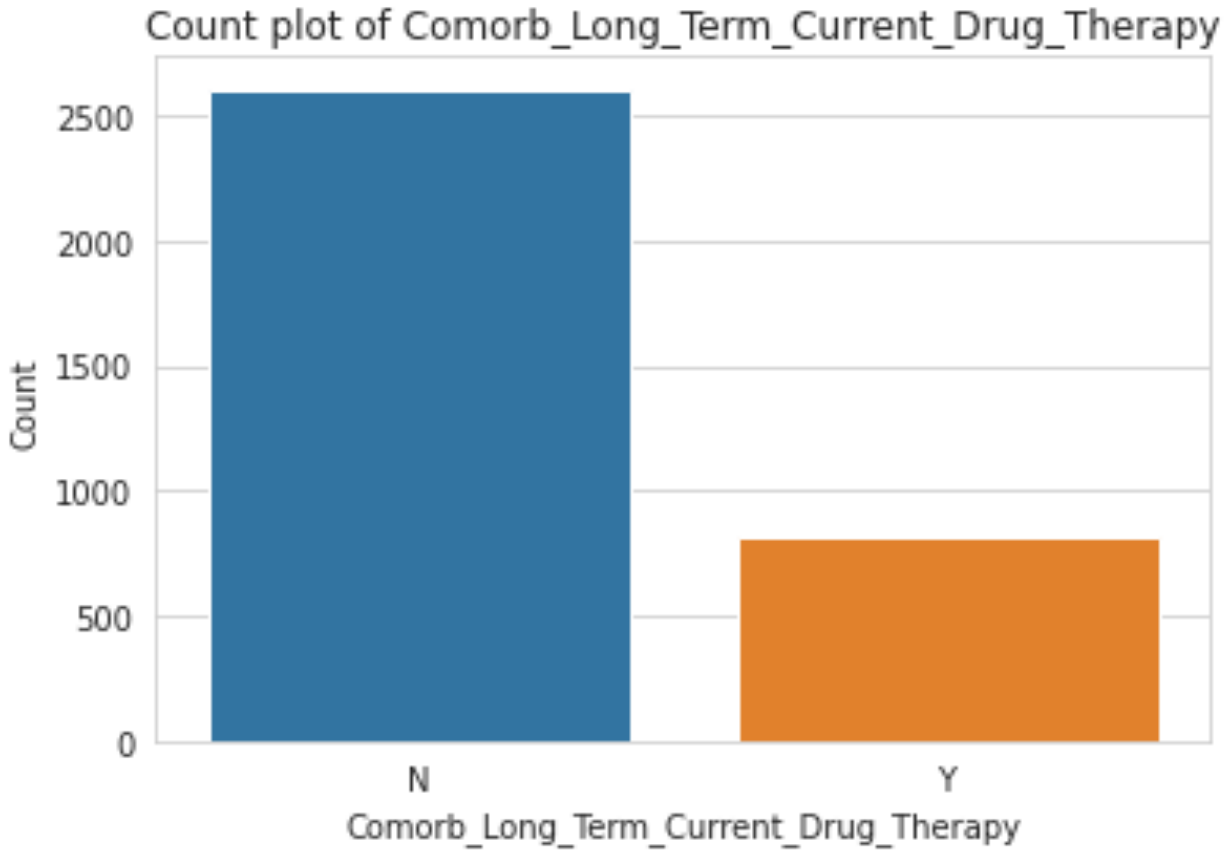
Class separation by categorical features

Using bar plots to visualize the separation of the target variable by the various categorical features.

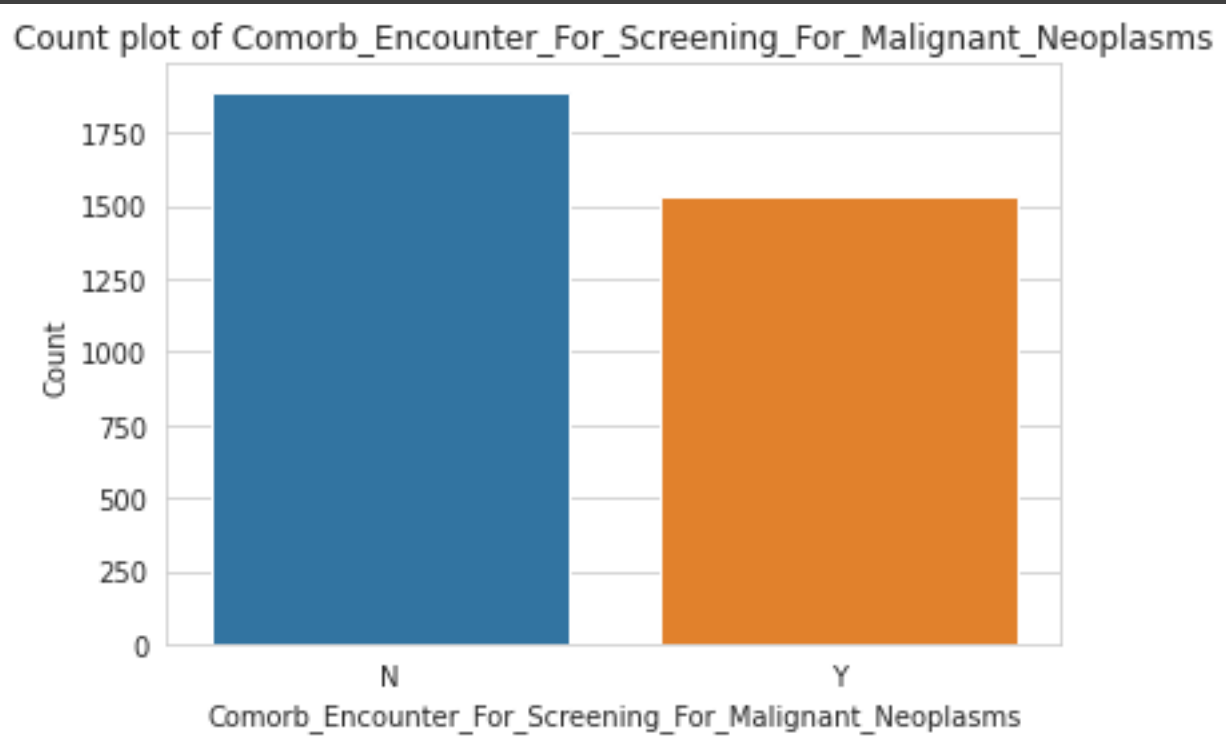
Class separation by Dexa_During_Rx



Class separation by Comorb_Long_Term_Current_Drug_Therapy

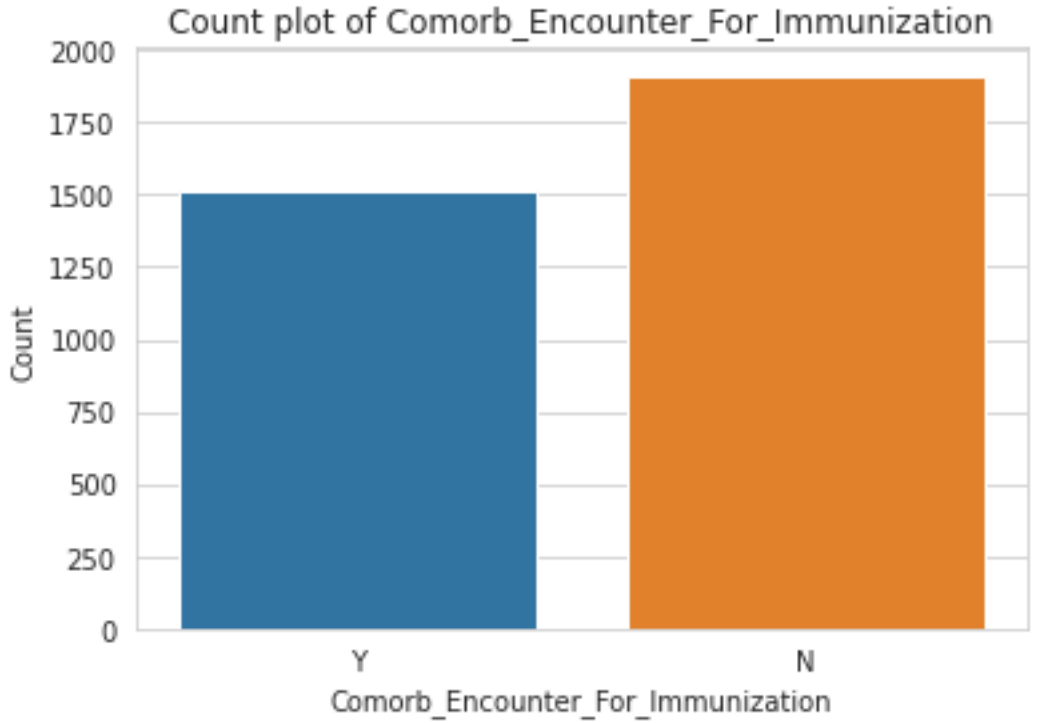


Class separation by Comorb_Encounter_For_Screening_For_Malignant_Neoplasms



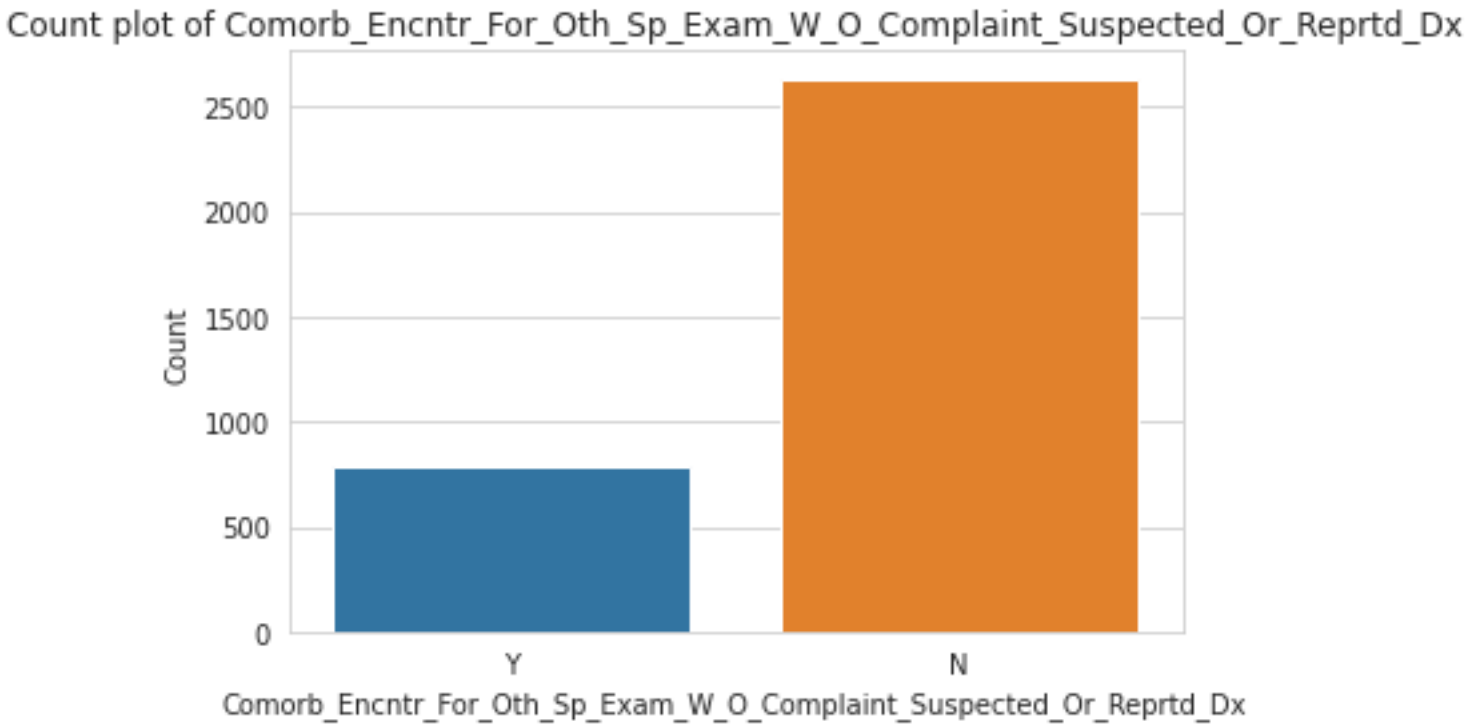
- It may be a misleading data compared to other results. The results are close to each other.

Class separation by Comorb_Encounter_For_Immunization

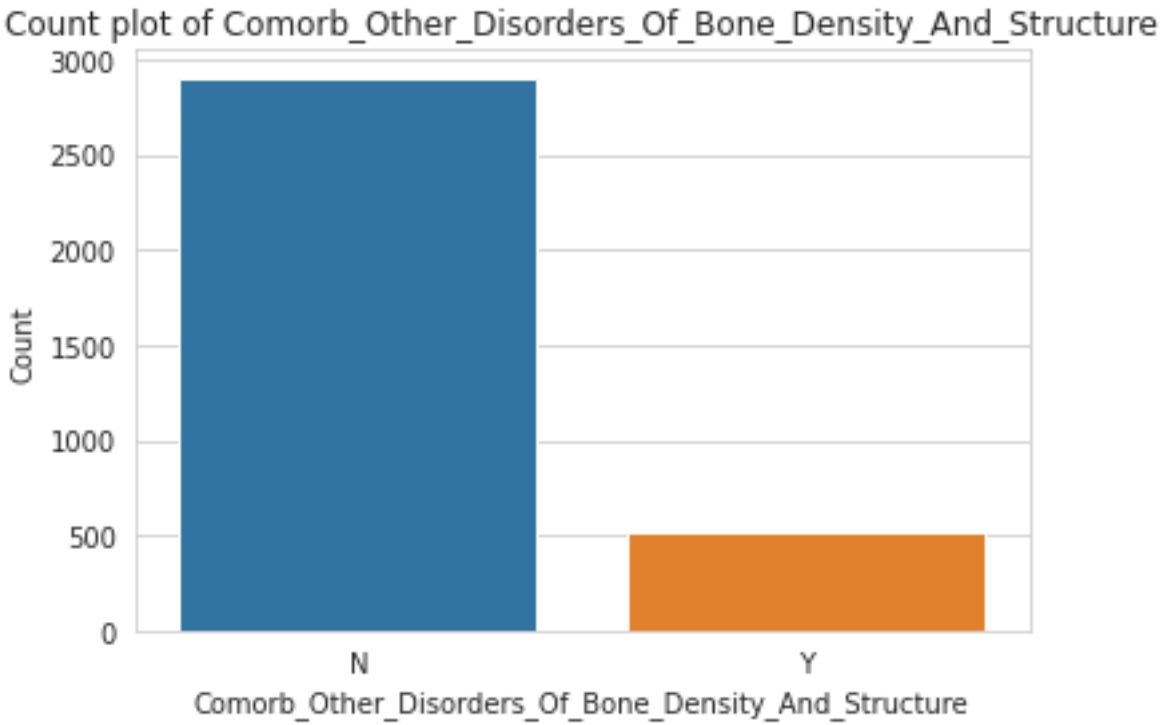


- The data are quite close.

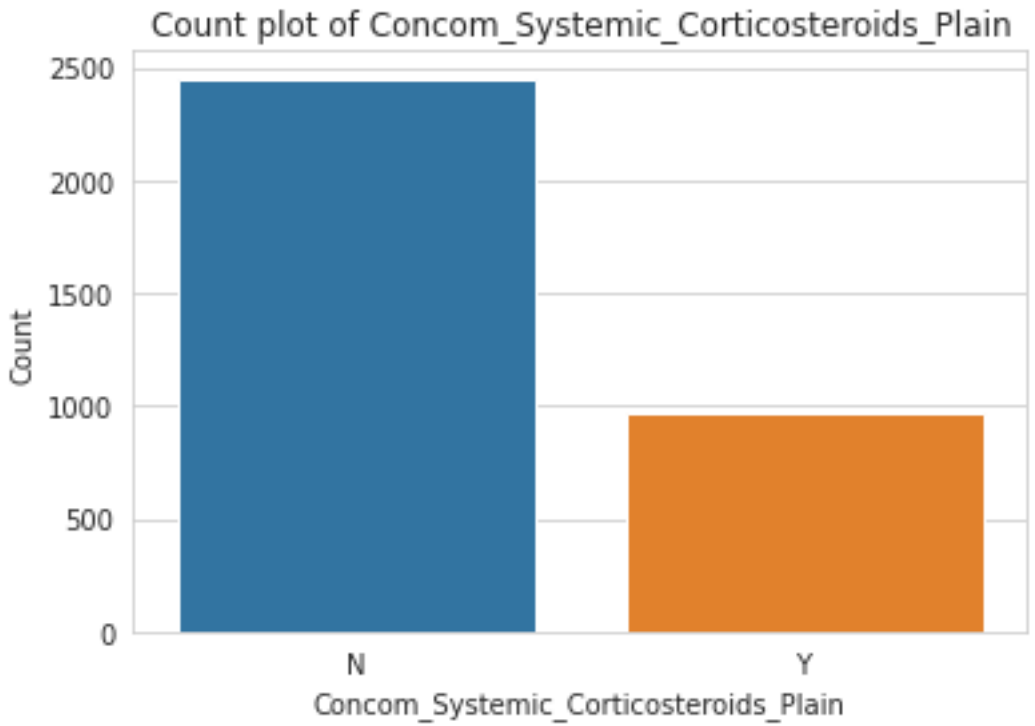
Class separation by Comorb_Encntr_For_General_Exam_W_O_Complaint, _Susp_Or_Reprtd_Dx



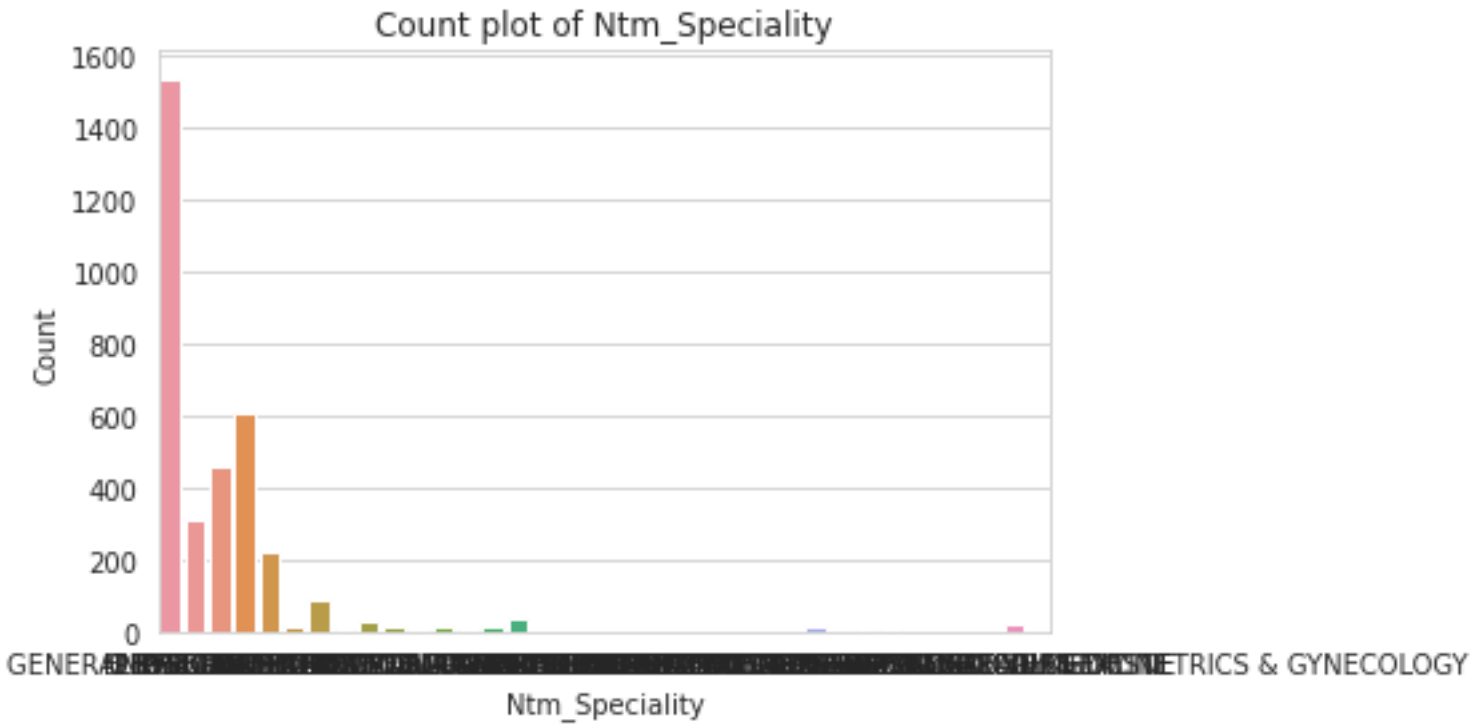
Class separation by Comorb_Other_Disorders_Of_Bone_Density_And_Structure



Class separation by Concom_Systemic_Corticosteroids_Plain

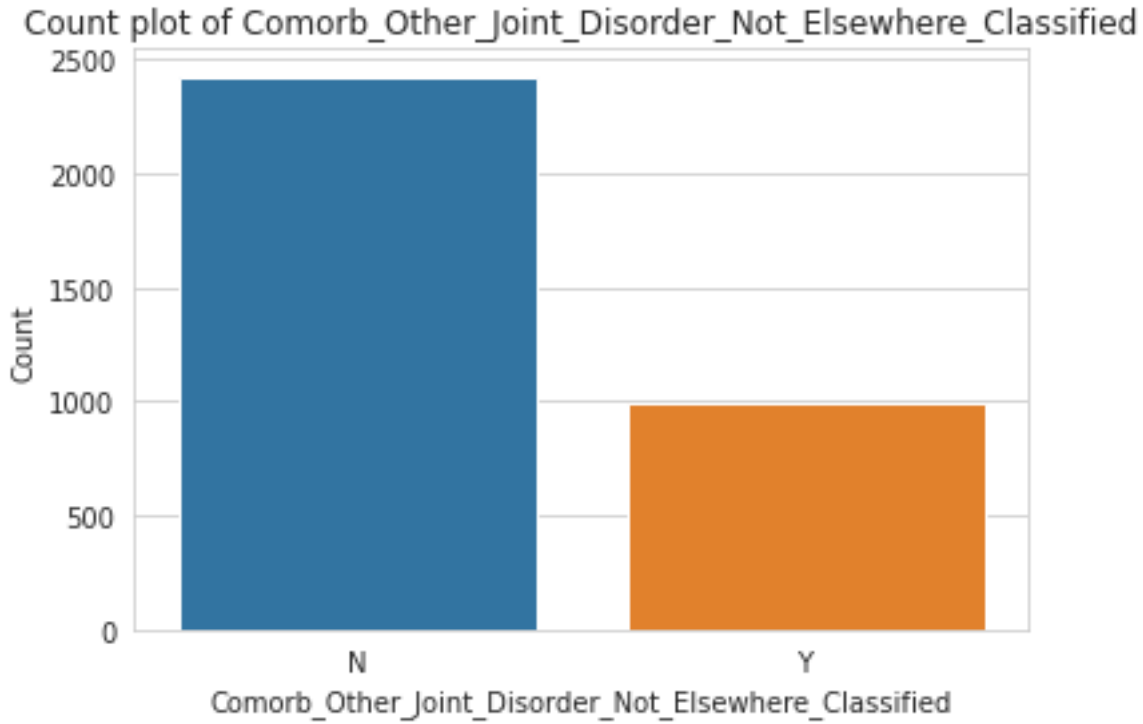


Class separation by Ntm_Speciality

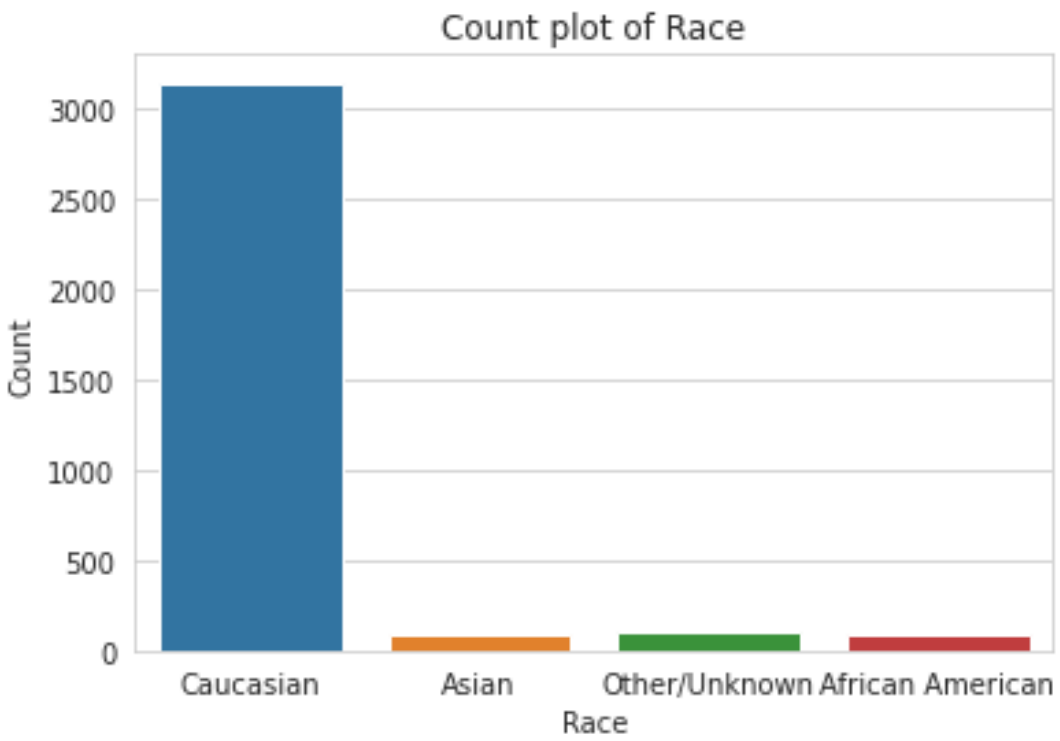


Although some of the data is spread over the whole, the majority is concentrated in one region.

Class separation by Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified

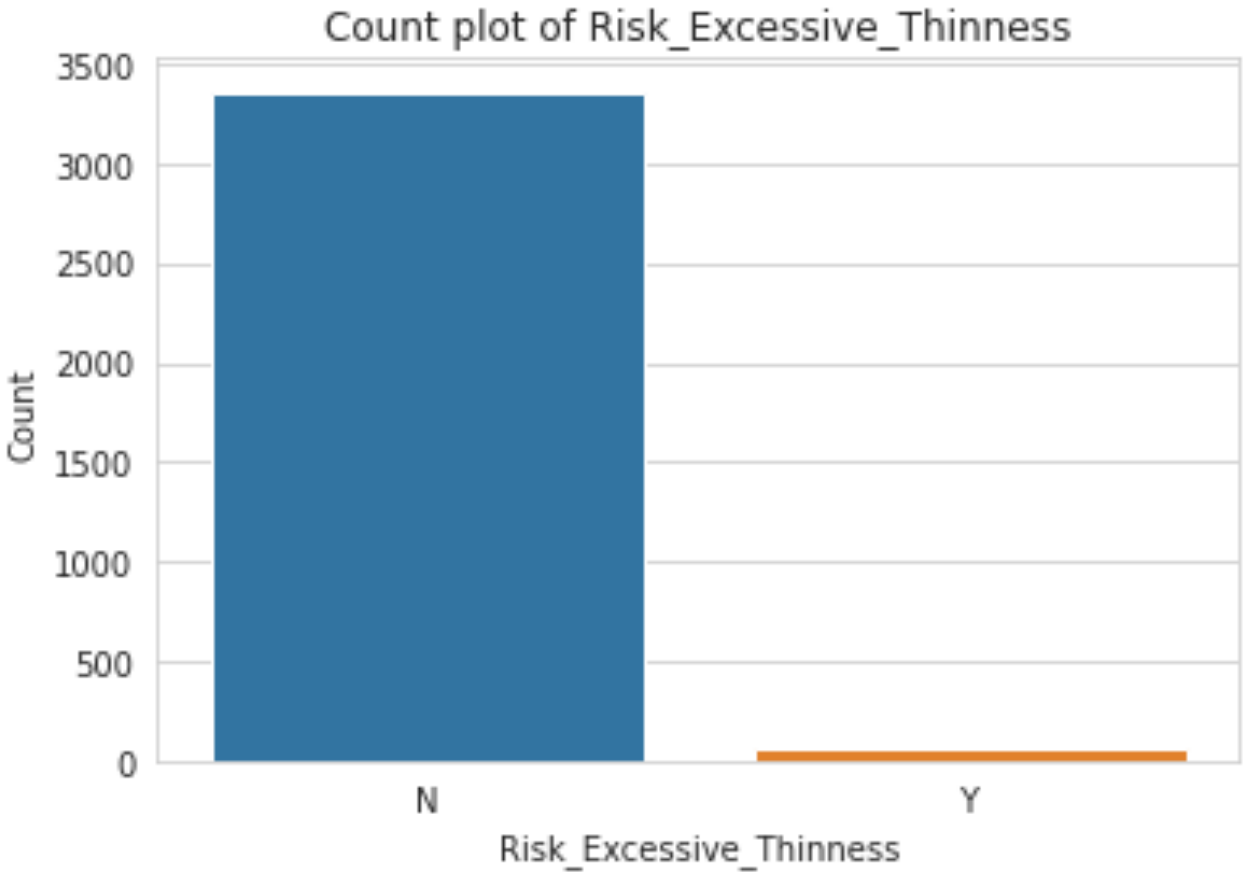


Class separation by Race

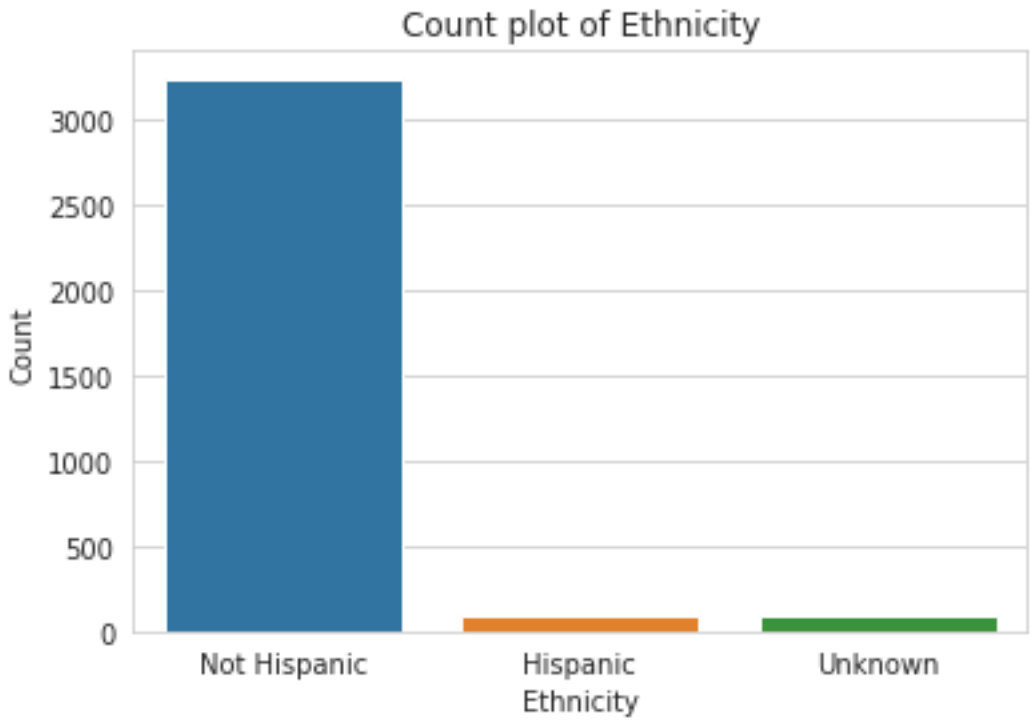


- The data is highly skewed.

Class separation by Risk_Excessive_Thinness



Class separation by Ethnicity



- The graph changed in 3 different factors.

EDA Summary

- From the Exploratory Data Analysis done, we are able to find how the different features/variables affects drug persistency.
- The categorical variables that have higher MI scores have a greater effect in the drug persistency as compared to those that have low MI scores.

Recommendations

For the purpose of automating the process of drug persistency identification, the following machine learning models can be used:

- **Logistic regression** – It is a type of linear model that is used for binary classification. It predicts output which is a categorical dependent variable. Such predictions are like yes or no, A or B, etc.
- **Decision tree**-They are good classifiers which are robust against outliers
- **LightGBM Classifier** – This is high-performance gradient boosting framework based on decision tree that is used for classification.

Thank You