Exploratory Data Analysis

ABC Pharma
Julia Donato

16-March-2023

Agenda

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Problem Statement

- ABC Pharma would like to gain insight on the persistency of their drug as per the physician's prescription.
- Objective: Perform an exploratory data analysis on a healthcare data set in order to aid in data understanding and in providing a recommendation for a predictive model.

EDA –Data Understanding

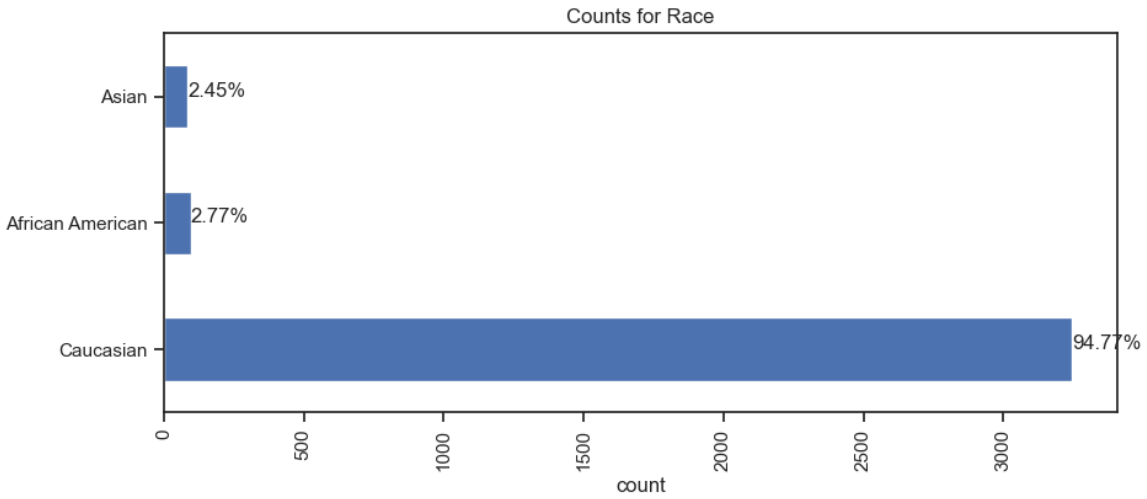
- Healthcare_dataset.xlsx cleaned, removing three rows, converted to cleaned_dataset.csv
- Information on 3,424 patients
- Data Types: Int64 and object
- Some object values converted to numeric (eg. ‘Y’ : ‘0’ , ‘N’: ’1’)

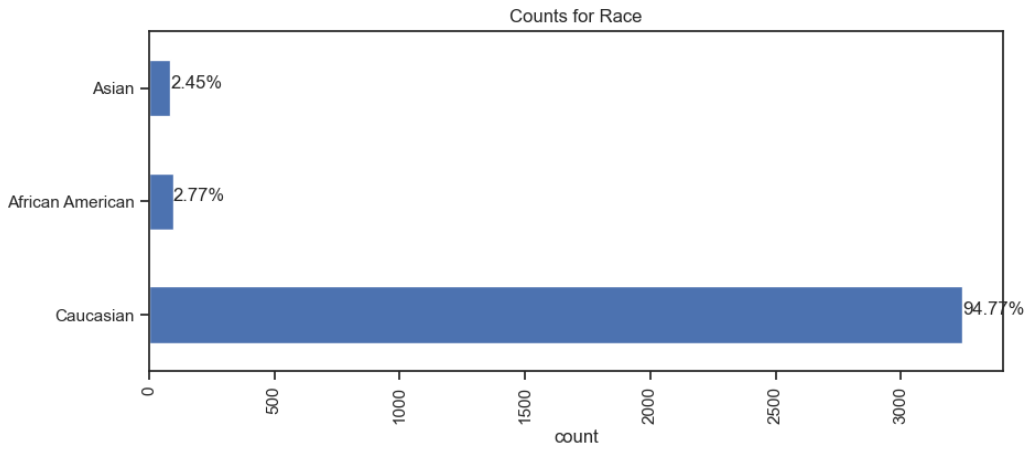
EDA – Categorical Features

	count	unique	top	freq
Ptid	3424	3424	P1	1
Persistency_Flag	3424	2	Non-Persistent	2135
Gender	3424	2	Female	3230
Race	3424	3	Caucasian	3245
Ethnicity	3424	2	Not Hispanic	3326
Region	3424	4	Midwest	1443
Ntm_Speciality	3424	36	GENERAL PRACTITIONER	1535
Ntm_Specialist_Flag	3424	2	Others	2013
Ntm_Speciality_Bucket	3424	3	OB/GYN/Others/PCP/Unknown	2104
Change_T_Score	3424	4	No change	1660
Adherent_Flag	3424	2	Adherent	3251

- Though majority non-persistent, still majority adherent (proportion of drug taken significant)
- Majority Caucasian
- Majority Midwest
- Majority female
- NTM Speciality majority General practitioner

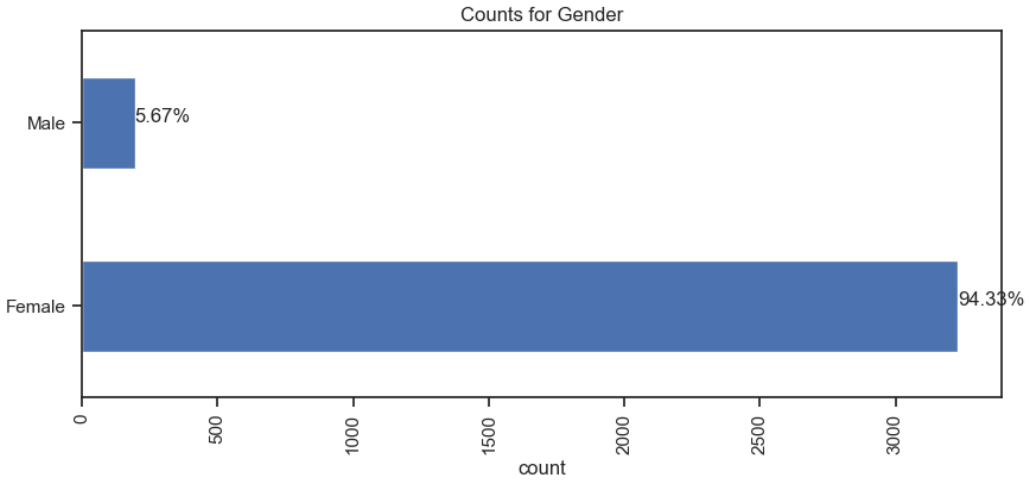
EDA – Categorical Features

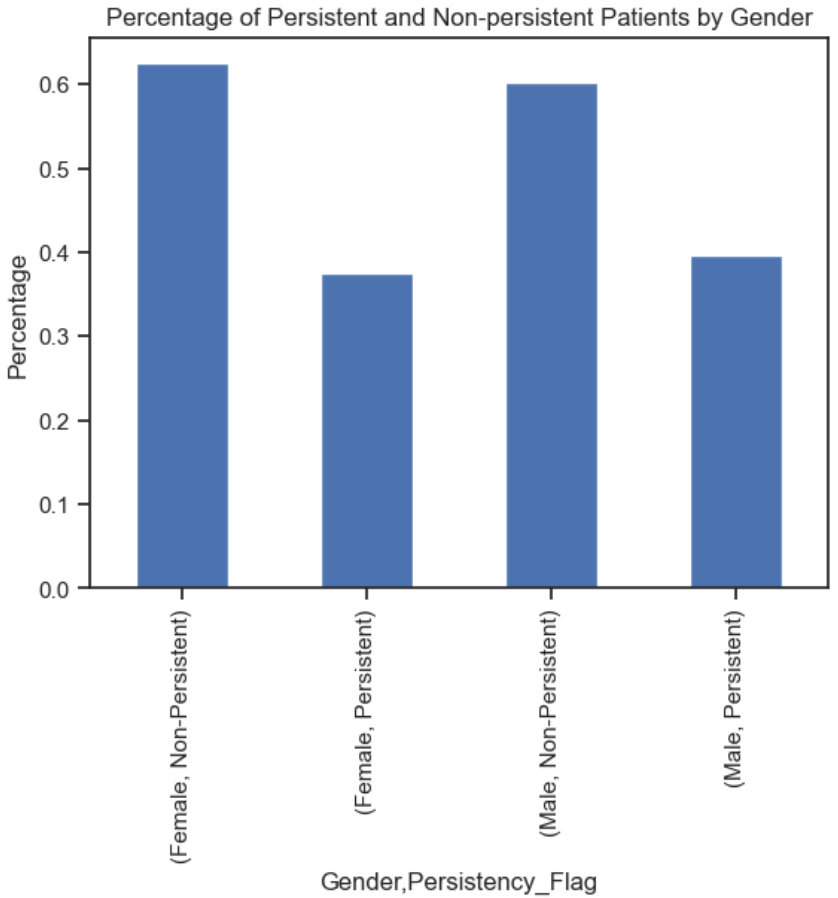




- It is likely that a combination of genetic, cultural and regional factors are at play each to some extent

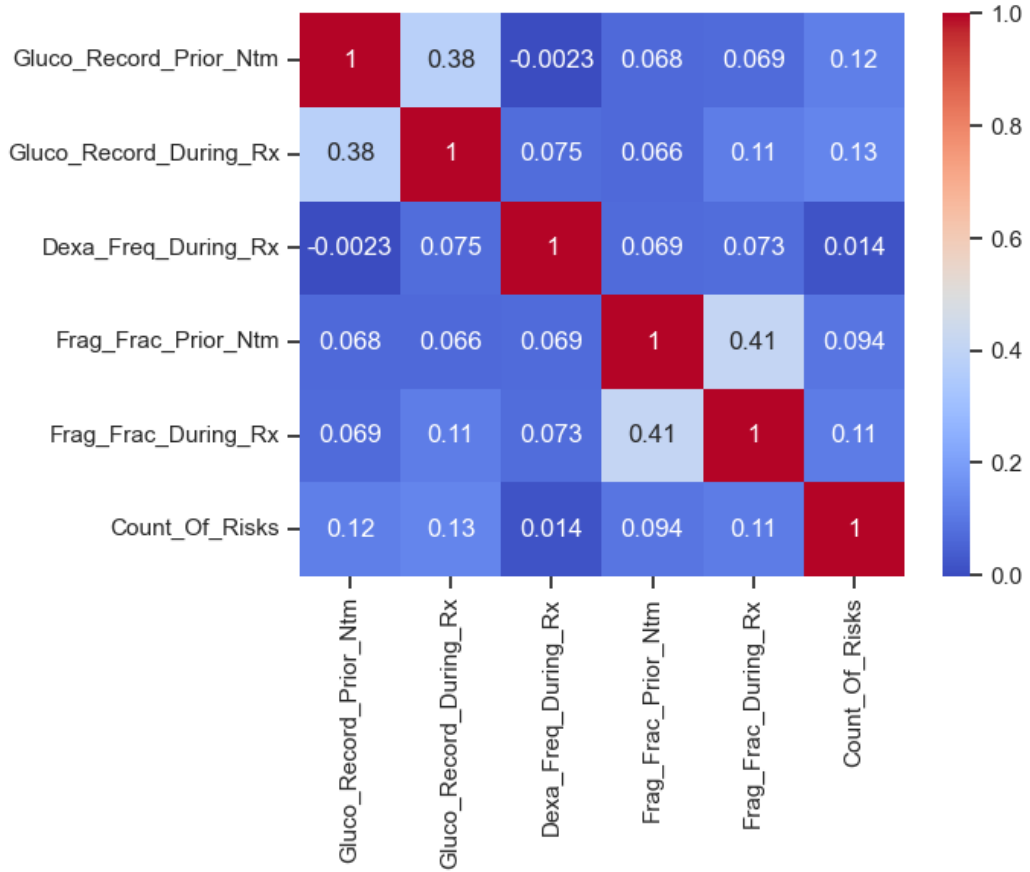
EDA – Categorical Features





- Women are much more likely to be affected than men
- Similar persistency rates among women and men

EDA – Correlations

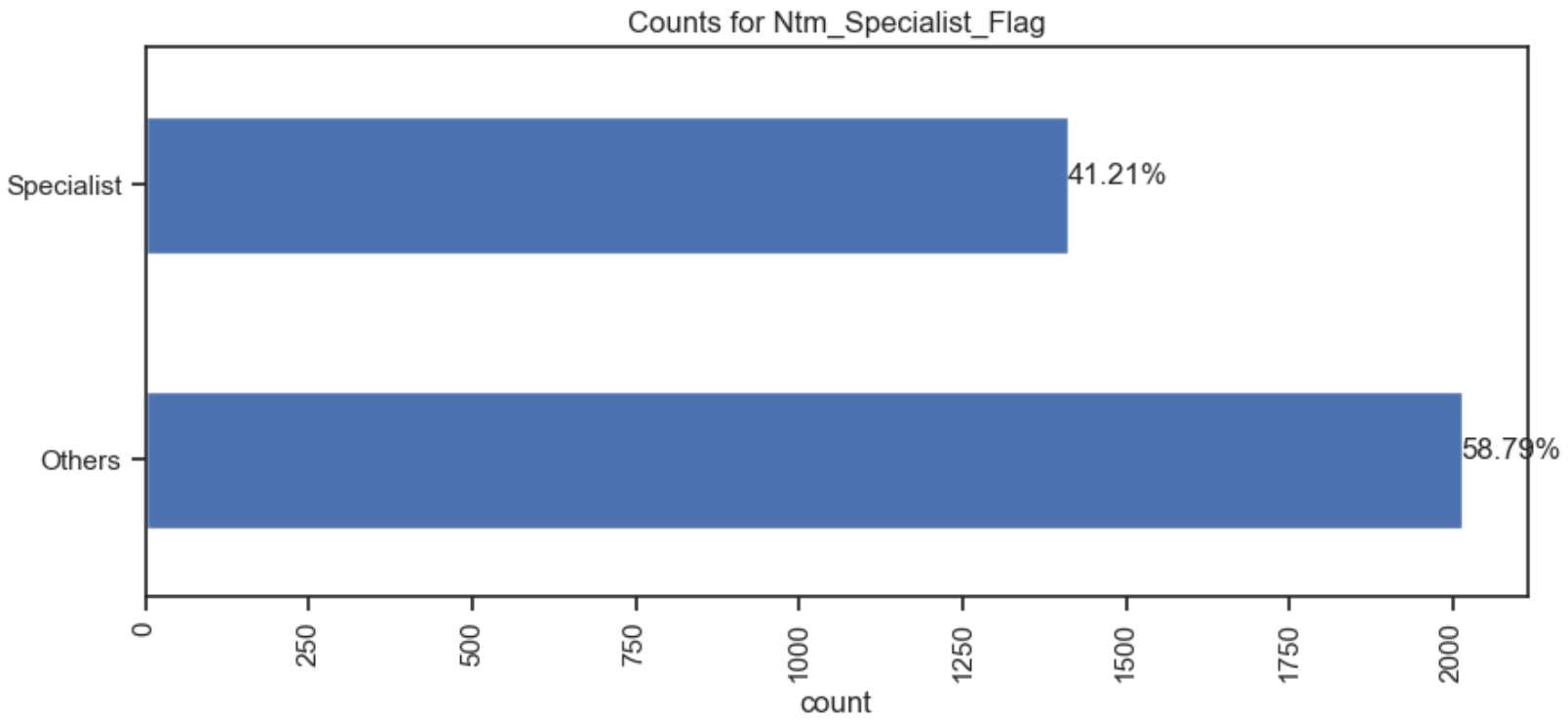


Persistency_Flag	Non-Persistent	Persistent
Gluco_Record_Prior_Ntm		
0	0.621993	0.378007
1	0.628571	0.371429

Persistency_Flag	Non-Persistent	Persistent
Gluco_Record_During_Rx		
0	0.68517	0.31483
1	0.45122	0.54878

- Significant correlation between fracture before and during treatment as well as Glucose
- Glucose levels during Rx show that this variable will be important in predictive models

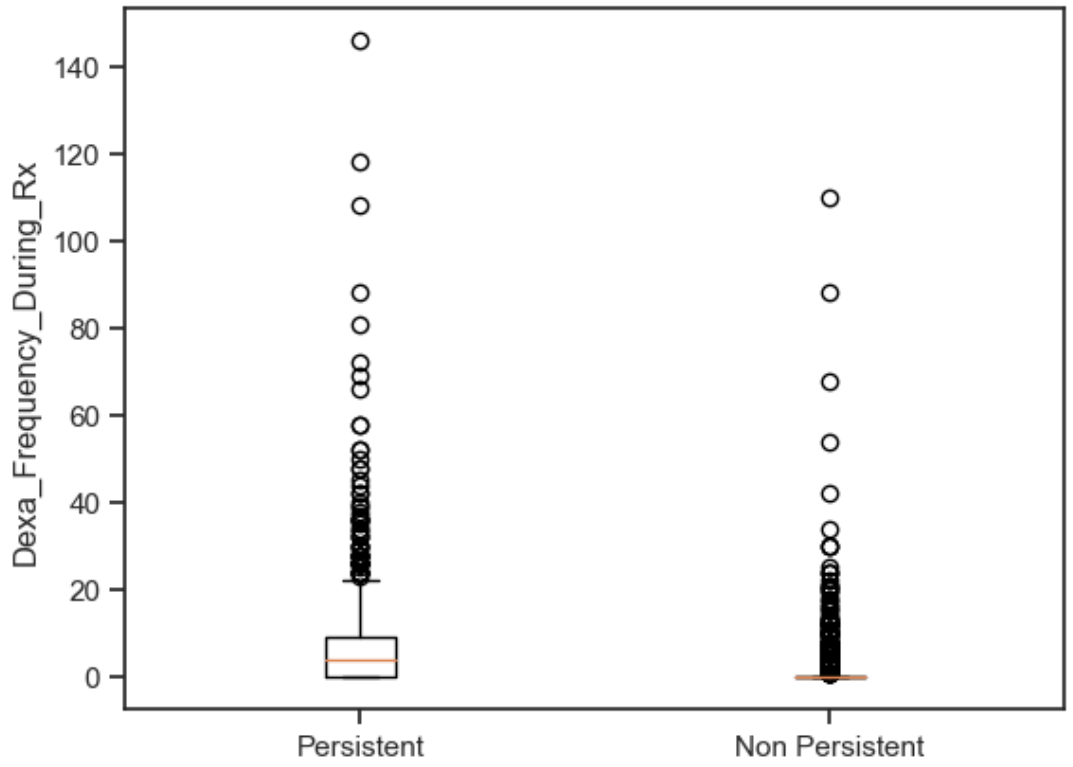
EDA



Persistency_Flag	Non-Persistent	Persistent
Ntm_Specialist_Flag		
Others	0.680079	0.319921
Specialist	0.542877	0.457123

- More persistency in those seeing a specialist but most patients are not

EDA



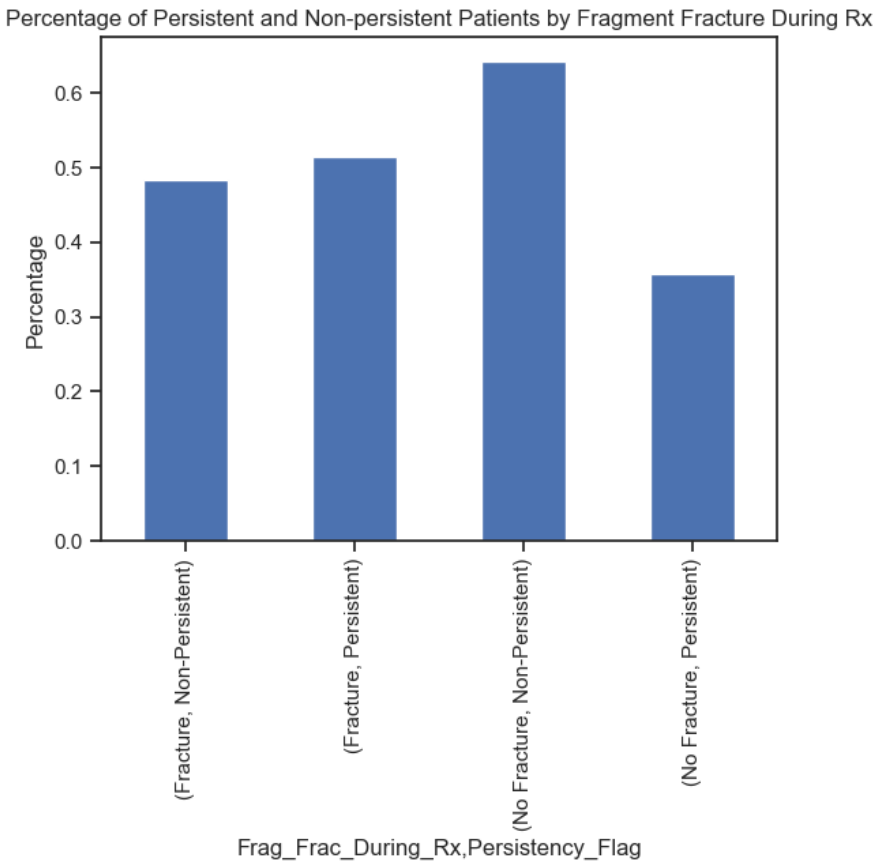
- Persistent patients more frequently have Dexa scans performed

EDA – Comorb/Risk Factors

Gender	Female	Male
	<lambda>	<lambda>
Injectable_Experience_During_Rx	0.891950	0.902062
Idn_Indicator	0.750464	0.685567
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias	0.517647	0.479381
Risk_Vitamin_D_Insufficiency	0.481734	0.412371
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	0.460991	0.226804
Comorb_Encounter_For_Immunization	0.441176	0.453608
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx	0.388854	0.494845
Concom_Narcotics	0.359133	0.376289
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations	0.344272	0.360825
Comorb_Vitamin_D_Deficiency	0.322910	0.257732
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	0.291950	0.288660
Concom_Systemic_Corticosteroids_Plain	0.284520	0.278351
Concom_Anti_Depressants_And_Mood_Stabilisers	0.281734	0.252577
Comorb_Osteoporosis_without_current_pathological_fracture	0.267802	0.268041
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx	0.237461	0.123711
Comorb_Long_Term_Current_Drug_Therapy	0.235604	0.288660
Comorb_Dorsalgia	0.224768	0.273196
Comorb_Personal_History_Of_Other_Diseases_And_Conditions	0.195975	0.226804
Risk_Smoking_Tobacco	0.187616	0.195876
Comorb_Personal_history_of_malignant_neoplasm	0.187307	0.226804
Concom_Fluoroquinolones	0.186687	0.175258
Comorb_Gastro_esophageal_reflux_disease	0.184211	0.180412
Concom_Cephalosporins	0.172136	0.242268
Concom_Macrolides_And_Similar_Types	0.165635	0.185567
Comorb_Other_Disorders_Of_Bone_Density_And_Structure	0.152941	0.123711
Concom_Anaesthetics_General	0.143963	0.164948
Risk_Chronic_Malnutrition_Or_Malabsorption	0.136223	0.154639
Concom_Broad_Spectrum_Penicillins	0.127864	0.134021
Risk_Family_History_Of_Osteoporosis	0.106192	0.077320
Concom_Viral_Vaccines	0.102167	0.118557
Risk_Patient_Parent_Fractured_Their_Hip	0.076471	0.046392
Risk_Poor_Health_Frailty	0.055728	0.061856
Risk_Rheumatoid_Arthritis	0.038700	0.025773
Risk_Type_1_Insulin_Dependent_Diabetes	0.037771	0.087629
Risk_Untreated_Chronic_Hypogonadism	0.032198	0.118557
Risk_Excessive_Thinness	0.019505	0.020619
Risk_Recurring_Falls	0.018885	0.041237
Risk_Hysterectomy_Oophorectomy	0.016718	0.000000
Risk_Low_Calcium_Intake	0.013003	0.000000
Risk_Chronic_Liver_Disease	0.005263	0.005155
Risk_Immobilization	0.004025	0.005155
Risk_Untreated_Early_Menopause	0.003715	0.000000
Risk_Estrogen_Deficiency	0.003406	0.000000
Risk_Osteogenesis_Imperfecta	0.000929	0.000000
Risk_Untreated_Chronic_Hyperthyroidism	0.000619	0.000000

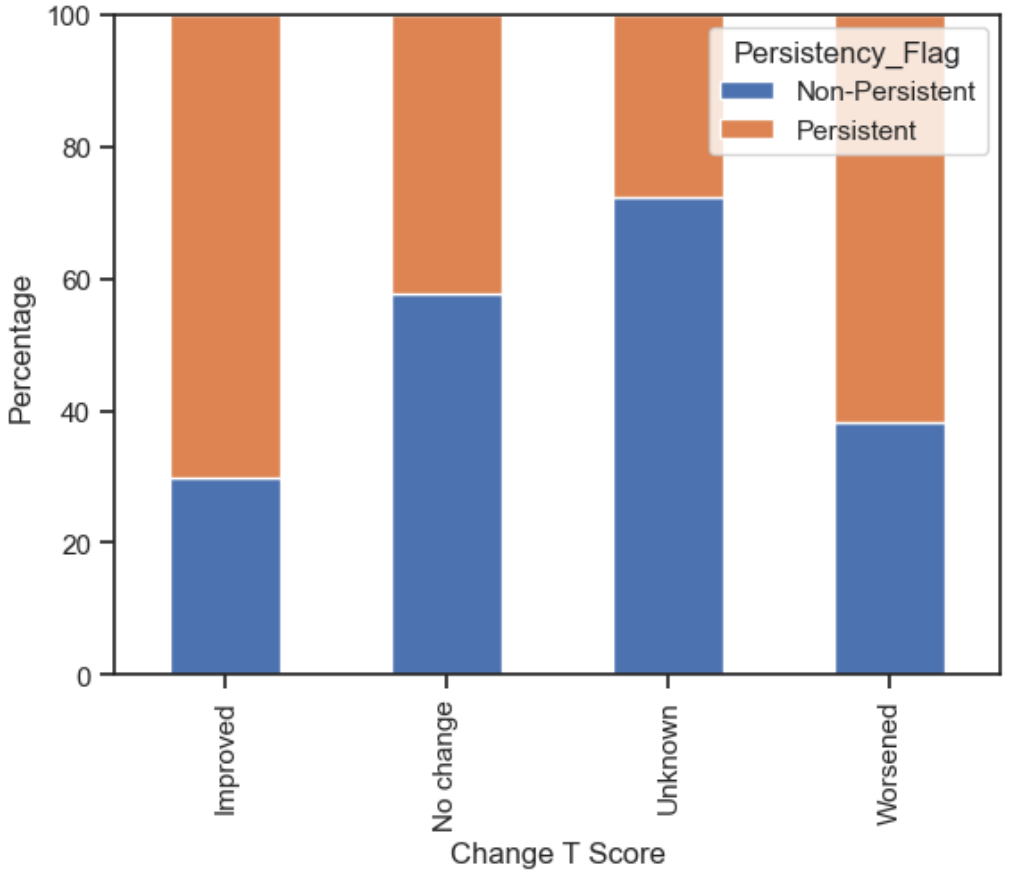
- Highest comorbidity factor related to lipoprotein and other lipidemia suggesting a correlation
- Patients highest risk factor is Vitamin D insufficiency more prominent in women
- Comorbidity factors more common than risk factors
- More than twice the amount of women have comorbidity with malignant neoplasms than men

EDA



- Persistent patients slightly more prone to fracture

EDA



- Persistent patients more likely to improve T Score than non persistent, however more worsen than non persistent
- May not be an good factor for prediction

Recommendations

- Based on the complexity of the healthcare dataset, I recommend exploring multiple models to improve the accuracy of predictions. This would involve selecting a base model and then exploring different types of models from various families, such as Linear models, Ensemble models, and Boosting models. By doing so, we can gain a better understanding of the data and the strengths and weaknesses of different algorithms, ultimately leading to better predictions and more informed investment decisions.
- Possible Models: Logistic Regression, Random Forest, AdaBoost

Thank You