

## Data Glacier Internship – Week 10

### Group Name: Persistency of my own

Name	E-mail	Country	College	Specialization
Nahari Terena	naharifterena@gmail.com	Brazil	La Sapienza University of Rome	Data Science

## Healthcare - Persistency of a drug

### Problem Description

One of the challenges for all pharmaceutical companies is to understand the persistence of drugs as per the physician's prescription. To solve this problem, ABC pharma company approached an analytics company to automate this process of identification.

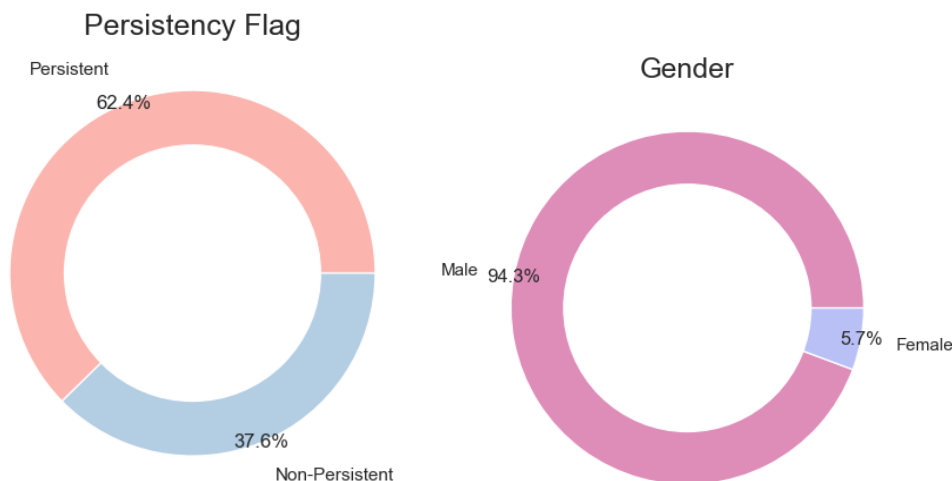
### Business Understanding

We aim to develop a web app for a pharmaceutical company to predict if a patient will get or not a drug schedule. Since we want to split the patients into categories considering their characteristics, we will develop a classification model.

### Exploratory Data Analysis

#### Demographics Unidimensional Analysis

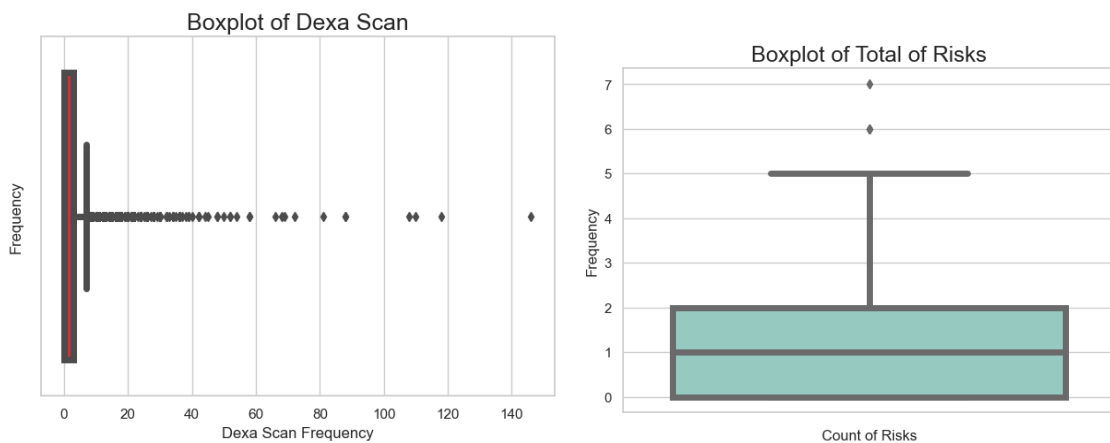
Over 60% of patients are persistency to a drug. Of all patients, 94.3% are male. Concerning race, Caucasian is the most common (91%), followed by Asian and Other/Unknown. Not Hispanic is 97% of the patients. Around 40% are from the West region. Only 4.8% is less than 55 years old.





## Outliers Analysis

We decided to maintain the outliers so we may probably identify these patients as a group itself.



## Main Questions

- How do risk factors relate to demographics?

Gender		Female	Male
		<lambda>	<lambda>
	Injectable_Experience_During_Rx	0.891950	0.902062
	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

There are some significant differences between genders:

- Women seem to be more affected by vitamin D deficiencies.
- More than twice as many women as men have passed as screening for malignant neoplasms.
- Four times as many men as women suffer from Hypogonadism (untreated).

There are some risks and other factors that seem to be significantly higher in South and West regions. It might be interesting to find out about socioeconomic factors aside.

Race		African American	Asian	Caucasian	Other/Unknown
		<lambda>	<lambda>	<lambda>	<lambda>
	Injectable_Experience_During_Rx	0.926316	0.880952	0.894854	0.793814
	Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	0.484211	0.464286	0.447268	0.412371
	Comorb_Encounter_For_Immunization	0.421053	0.607143	0.438691	0.422680
	Comorb_Encntr_For_General_Exam_W_O_Complaint_Susp_Or_Reprtd_Dx	0.336842	0.571429	0.391995	0.391753
	Comorb_Vitamin_D_Deficiency	0.463158	0.297619	0.314485	0.350515
	Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	0.294737	0.380952	0.286531	0.381443
	Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx	0.242105	0.309524	0.228717	0.226804
	Comorb_Long_Term_Current_Drug_Therapy	0.210526	0.202381	0.240470	0.237113
	Comorb_Dorsalgia	0.178947	0.261905	0.229034	0.195876
	Comorb_Personal_History_Of_Other_Diseases_And_Conditions	0.157895	0.226190	0.198221	0.195876

There seem to be some remarkable differences between Asian and other races. They are probably due to cultural factors and other behaviors.

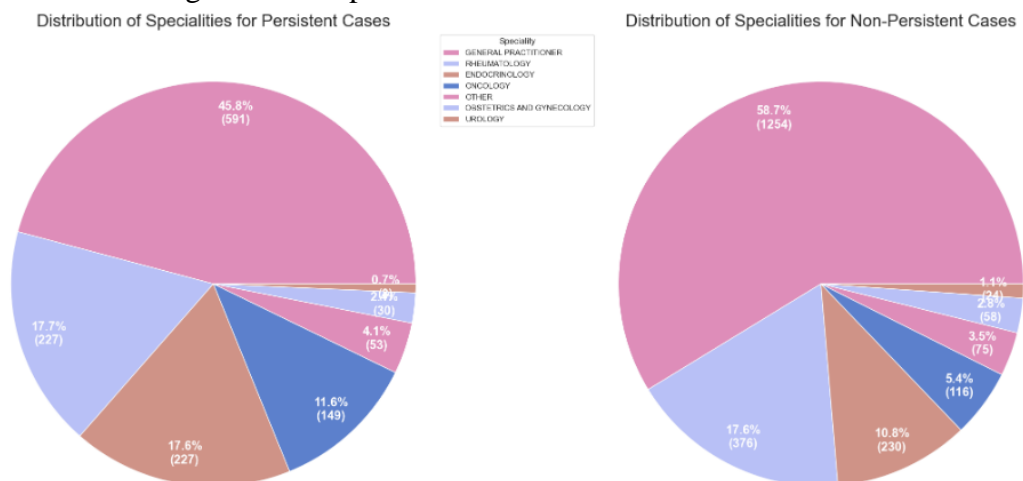
Age_Bucket		55-65	65-75	<55	>75
	Injectable_Experience_During_Rx	656	965	150	1285
	Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	361	595	77	500
	Comorb_Encounter_For_Immunization	257	517	46	693
	Comorb_Encntr_For_General_Exam_W_O_Complaint_Susp_Or_Reprtd_Dx	304	447	50	551
	Comorb_Vitamin_D_Deficiency	238	360	52	443
	Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	187	349	40	423
	Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx	226	269	40	256
	Comorb_Long_Term_Current_Drug_Therapy	160	254	29	374
	Comorb_Dorsalgia	131	248	36	364
	Comorb_Personal_History_Of_Other_Diseases_And_Conditions	128	216	32	301

Patients older than 65 years old are affect by the risks in higher proportion.

- What is the percentage of patients holding at least one factor?  
About 99% holds at least one risk, comorbid or concomitant factor.
- What are the most common risk factors?  
It's easy to see that most of the patients already hold comorbidity factors, while holding risk factors is less common.  
The main comorbidity factor is related to lipoproteins and metabolism (cholesterol). On the other hand, the main risk factor is deficiency in vitamin D. More than one third has been found to have taken narcotics.

Injectable_Experience_During_Rx	0.892523
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias	0.515479
Risk_Vitamin_D_Insufficiency	0.477804
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	0.447722
Comorb_Encounter_For_Immunization	0.441881
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx	0.394860
Concom_Narcotics	0.360105
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations	0.345210
Comorb_Vitamin_D_Deficiency	0.319217
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	0.291764
Concom_Systemic_Corticosteroids_Plain	0.284171

- Does “Ntm\_Speciality” variables seems to have useful information for the classification task?  
The distributions of frequency for the target variable by speciality are pretty similar. Thus, we may rule out the possibility that one of the factors is who prescribed the drug in the first place.



## Final Recommendation

Demographics variables seems to be relevant because it brings the context in which the patient is inserted. As well as comorbid and risks factors.

Regarding the prediction model, we may consider interpretable model such as decision tree to support the predictions of more complex models.

Github Repo <https://github.com/naharift/DataGlacier/tree/main/Week10>