Group Name: Attack on Data

Team Members:

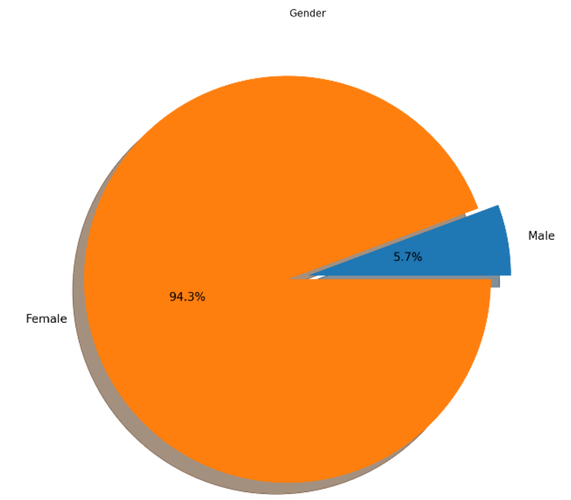
Name	Email	Country	College/Company	Specialization
Armin Khayati	Northatlas@gmail.com	UAE	-	Data Science
Ezzuldin Zaky	Ezzulding.zaky@gmail.com	UAE	American University of Sharjah	Data Science
Orcun Sami Tandogan	tandogan.orcun@metu.edu.tr	Turkey	Middle East Technical University	Data Science

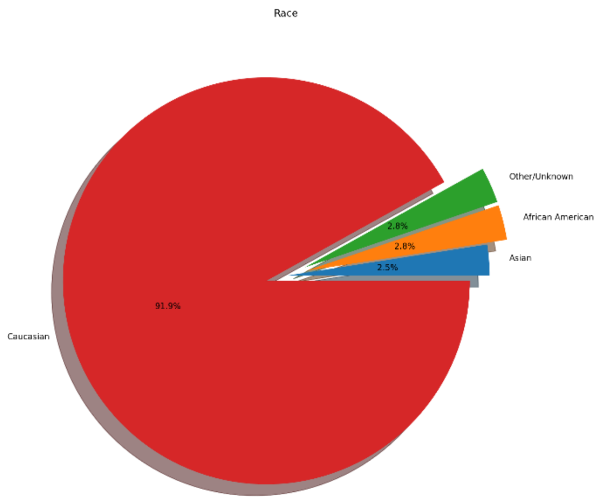
Problem description

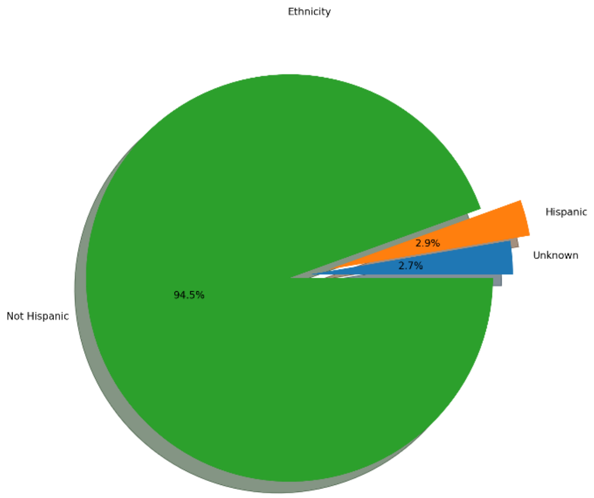
One of the challenges for all pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

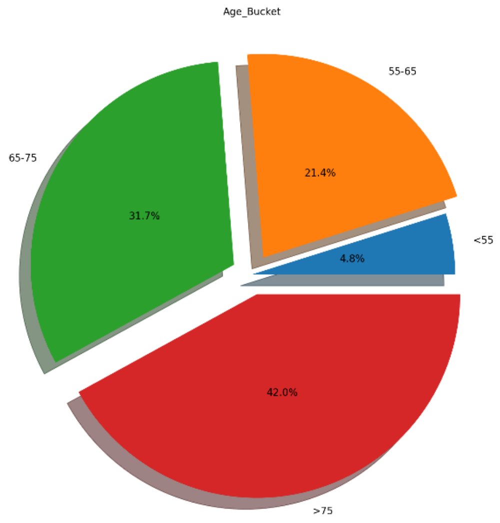
Exploratory data analysis (EDA)

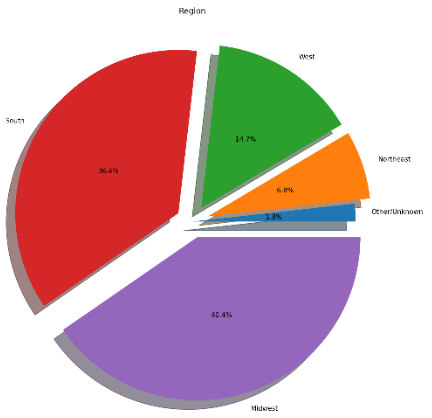
Our dataset has 69 columns, 67 of them are categorical and two of them are numerical. There are no nan values in the dataset. The main problem of it is that some columns have rare labels. As we can see in the bellow charts rare categories is the main issue.

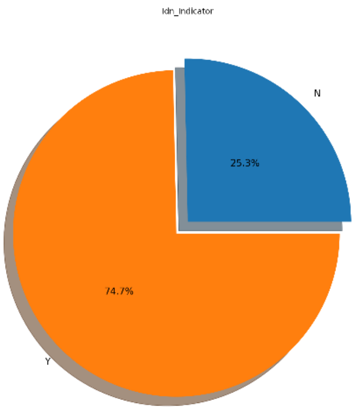

Gender
Male
5.7%
94.3%
Female

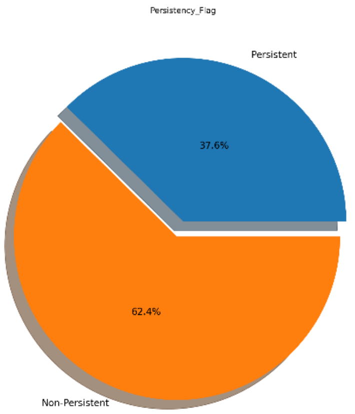

Race
Other/Unknown
2.8%
African American
2.8%
Asian
2.5%
91.9%
Caucasian

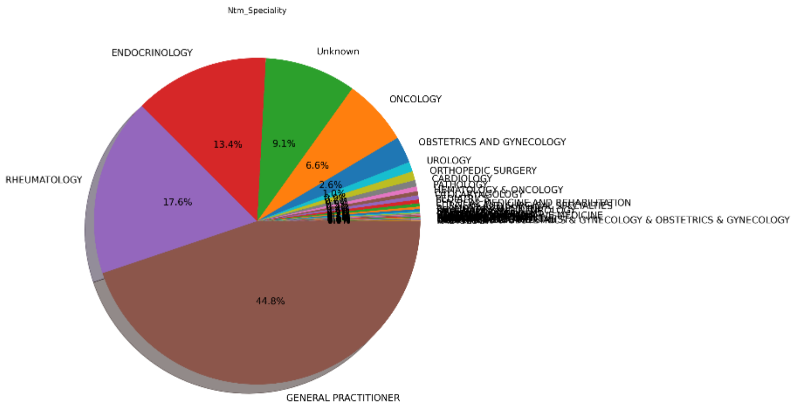

Ethnicity
Hispanic
2.9%
Unknown
2.7%
94.5%
Not Hispanic


Age_Bucket
55-65
65-75
21.4%
31.7%
<55
4.8%
42.0%
>75

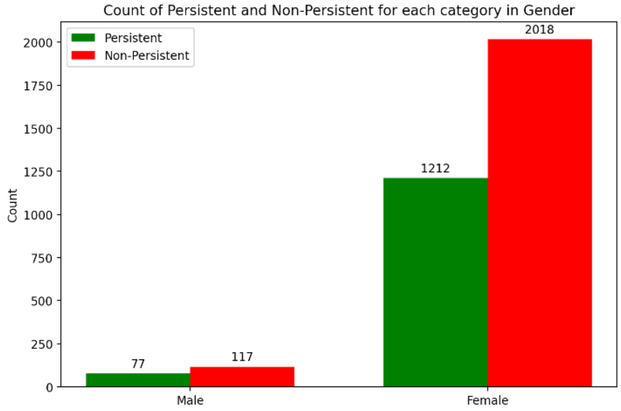


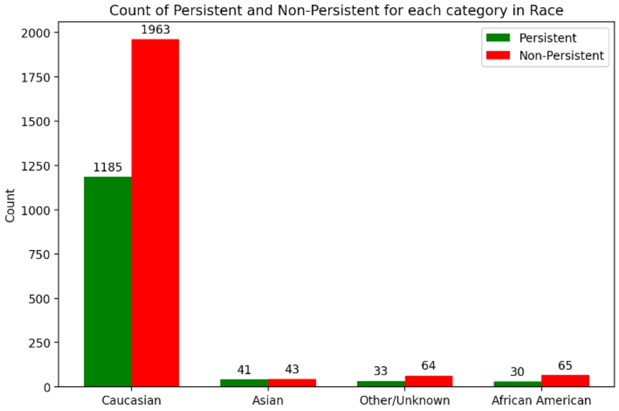


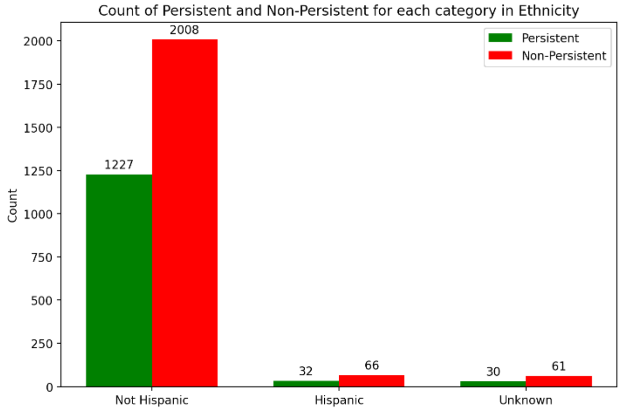


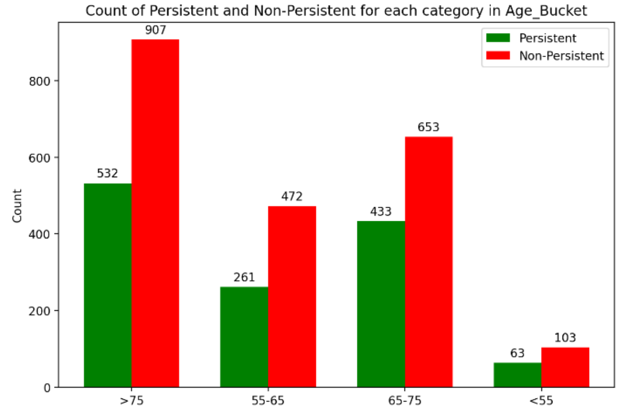


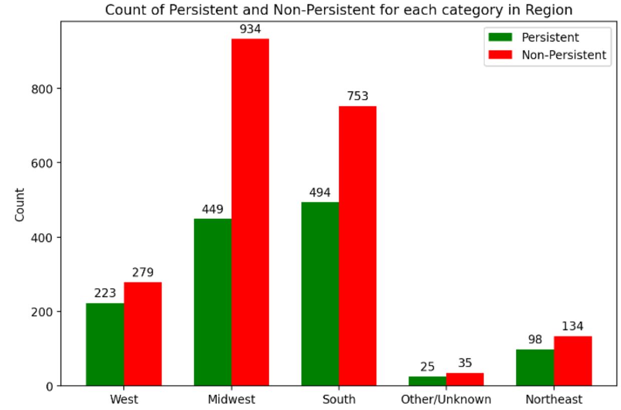
For gender variable 94% of values are female and only 5% are male and it means results for this dataset are mostly accurate for women. It is also the same for Caucasian Race and Not Hispanic Ethnicity. Also we can observe the imbalanceness of age_bucket variable for younger patients. It can be seen than most people on the dataset belonged to the aged class, but the non-persistency level (or ratio) is more among the older patients. As discussed earlier, this study has been imbalanced towards female subjects but among females, the non-persistency level is higher than the males. But no concrete conclusion can be drawn due to the data imbalance. Also, low risk patients were found to be less persistent than the high-risk ones.

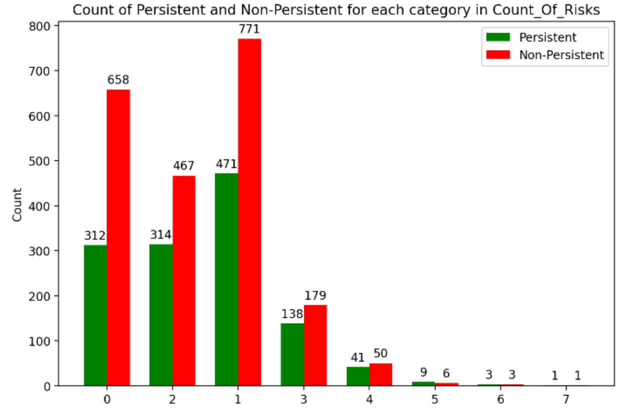

Count of Persistent and Non-Persistent for each category in Gender


Count of Persistent and Non-Persistent for each category in Race

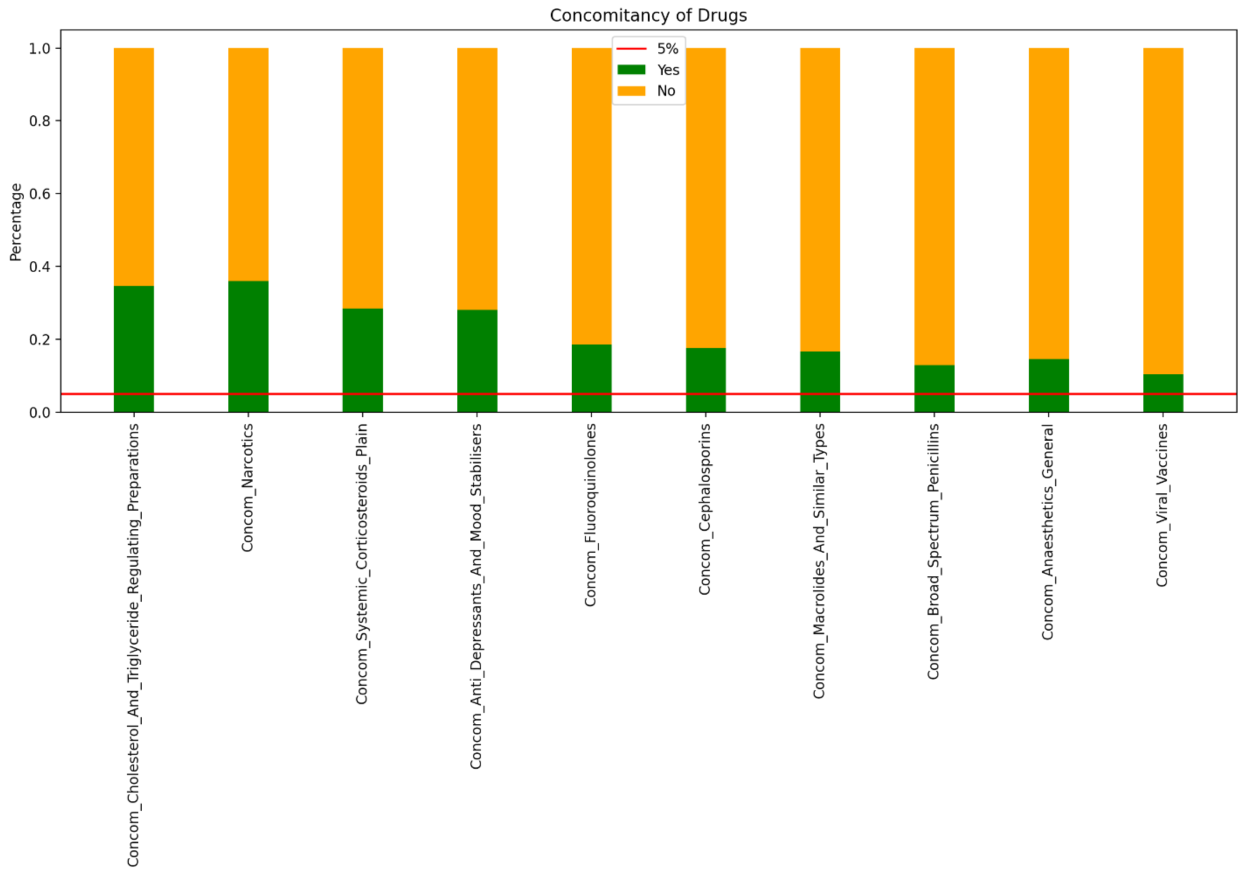

Count of Persistent and Non-Persistent for each category in Ethnicity


Count of Persistent and Non-Persistent for each category in Age_Bucket

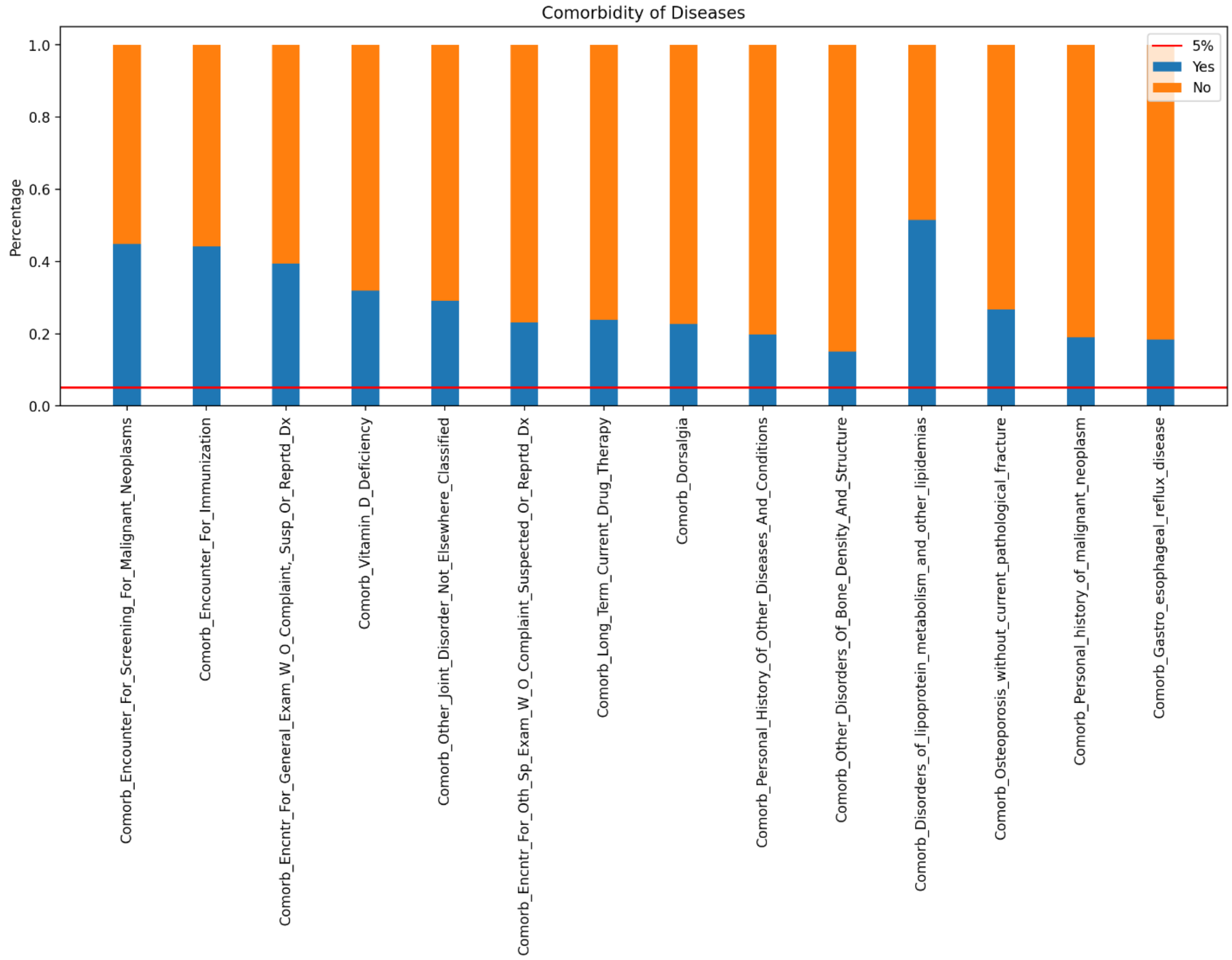

Count of Persistent and Non-Persistent for each category in Region


Count of Persistent and Non-Persistent for each category in Count_Of_Risks

By a look at Concomitancy of Drugs bellow we can see that Cholesterol drug is more used than others beside the main drug. For other drugs, Concomitancy percentage is very low.



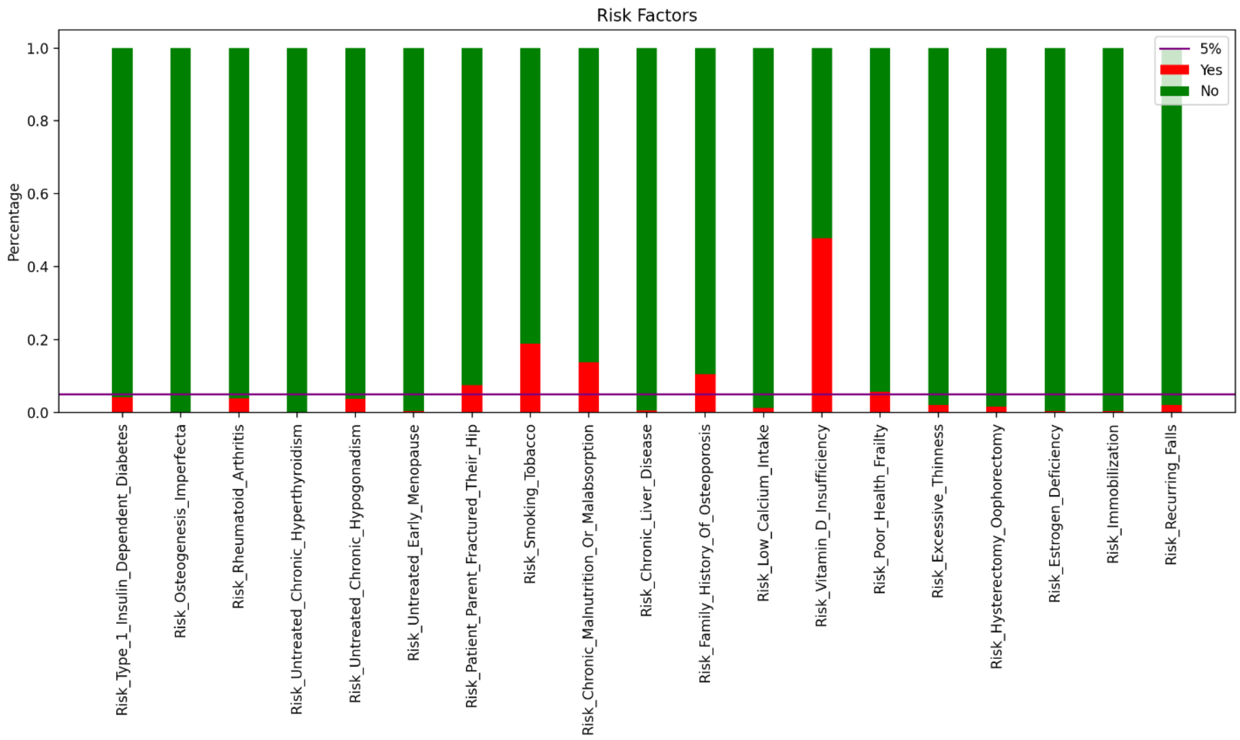
Also for Comorbidity of Diseases, we can see that disorders of lipoprotein have higher percentage to comorbid with main disease than other diseases and disorders of bone density is the lowest among all.



And among the risk factors most of them have less than 5% chance to endanger treatment. The risk factor with highest chance is Vitamin D Insufficiency and others above 5% are:

- Poor Health Frailty
- Family History Of Osteoporosis
- Chronic Malnutrition Or Malabsorption
- Smoking Tobacco
- Patient Parent Fractured Their Hip

Rest of the factors have less than 5% risk to endanger treatment.



Recommendation

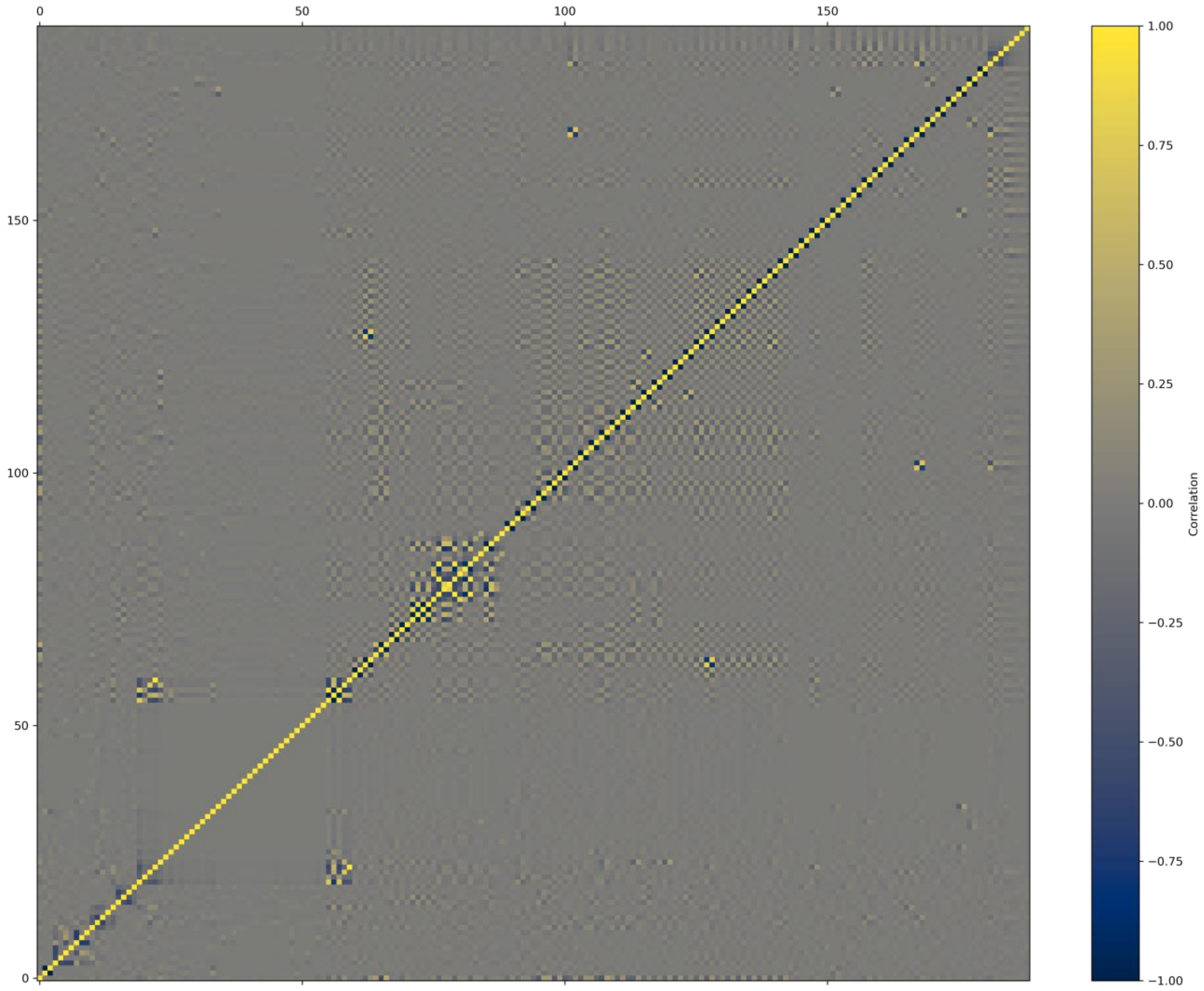
1. Handling **Unknown** values for Race, Region, and Ethnicity Variables
   - Using mode as an imputer as an imputer on ***Race*** and ***Ethnicity*** variables.
   - For ***Region*** variable, because most of the people with **Unknown** Region have **Not Hispanic** Ethnicity, and Most of people with Not Hispanic Ethnicity, have Midwest Region, we will replace Unknown Regions with **Midwest**.
2. Handling **Rare Labels**: Finding categories less than 5 percent in each variable, then merging those categories into one or drop them if the variable only has 2 categories (e.g., Y/N) and cardinality of one them is less than 5 percent.
3. Grouping integer values of Count_Of_Risks variable into two **bins**.

Group	Variable	Categories to Merge
	Race	African American, Asian
	Age_Bucket	<55, 55-65

Variables Chosen to Merge Categories	Ntm_Speciality	OBSTETRICS AND GYNECOLOGY , UROLOGY , ORTHOPEDIC SURGERY , CARDIOLOGY , PATHOLOGY , HEMATOLOGY & ONCOLOGY , OTOLARYNGOLOGY , PEDIATRICS , PHYSICAL MEDICINE AND REHABILITATION , PULMONARY MEDICINE , SURGERY AND SURGICAL SPECIALTIES , PSYCHIATRY AND NEUROLOGY , NEPHROLOGY , ORTHOPEDICS , PLASTIC SURGERY , VASCULAR SURGERY , HOSPICE AND PALLIATIVE MEDICINE , GERIATRIC MEDICINE , GASTROENTEROLOGY , TRANSPLANT SURGERY , CLINICAL NURSE SPECIALIST , OCCUPATIONAL MEDICINE , HOSPITAL MEDICINE , OPHTHALMOLOGY ,

		PODIATRY , EMERGENCY MEDICINE , RADIOLOGY , OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY , NEUROLOGY , PAIN MEDICINE , NUCLEAR MEDICINE
	Change_T_Score	Improved, Worsened
	Change_Risk_Segment	Improved, Worsened
	Count_Of_Risks	Bin 1 is [0,1,2,3] and bin 2 is [4,5,6,7]
Variables Chosen to Drop	Ethnicity	
	Risk_Type_1_Insulin_Dependent_Diabetes	
	Risk_Osteogenesis_Imperfecta	
	Risk_Rheumatoid_Arthritis	
	Risk_Untreated_Chronic_Hyperthyroidism	
	Risk_Untreated_Chronic_Hypogonadism	
	Risk_Untreated_Early_Menopause	
	Risk_Chronic_Liver_Disease	
	Risk_Low_Calcium_Intake	
	Risk_Excessive_Thinness	
	Risk_Hysterectomy_Oophorectomy	
	Risk_Estrogen_Deficiency	
	Risk_Immobilization	
	Risk_Recurring_Falls	
	Dexa_Freq_During_Rx	

4. **One hot encoding** all the variables after doing above tasks
5. Removing variables with more than **98% correlation**.



6. Try different machine learning approaches and select the best model.

Code Review

Code By	Review By
Armin Khayati	Ezzuldin Zaky
Ezzuldin Zaky	Orcun Sami Tandogan
Orcun Sami Tandogan	Armin Khayati

Github Repo:

https://github.com/Arminkhayati/dataglacier_internship