



Data Glacier

Your Deep Learning Partner

Drug Persistency

Presented by Emily Yao

(yaowemily@gmail.com, University of Florida, Data Science)

on 1/10/2023

Agenda

Background.....	3
Data Understanding.....	4
Quantitative Data Analysis.....	5-6
Patient Demographic.....	7-12
Categorical Data Analysis....	13-14
Recommendations (For ML).....	15
Machine Learning Results...	16-19
Sources.....	20





Background

Problem Statement:

- One challenge all Pharmaceutical companies face is to understand the persistency of a drug as per physician prescription. To solve this problem, ABC pharma company approached an analytics company to automate this process of identification.

ML Problem:

- With an objective to gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

Analysis (broken into five parts):

- Data understanding
 - Quantitative data analysis
 - Patient Demographic (categorical data)
 - Categorical data analysis
 - Recommendations for a classification machine learning algorithm
-

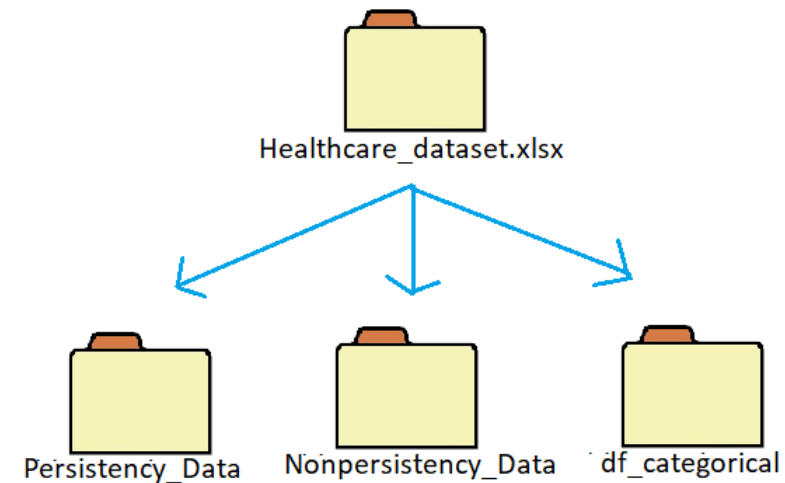
Data Understanding

General Characteristics:

- 69 features (columns)
- No specified time frame in dataset
- 3,424 patients
- Total data points: 236,256

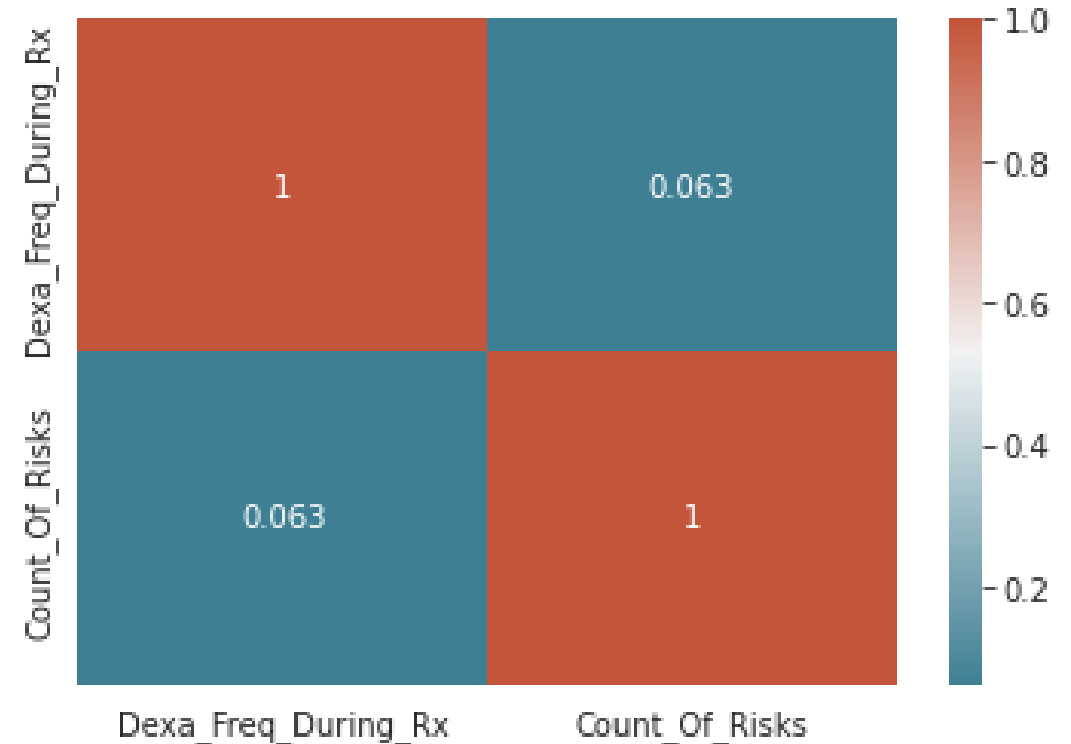
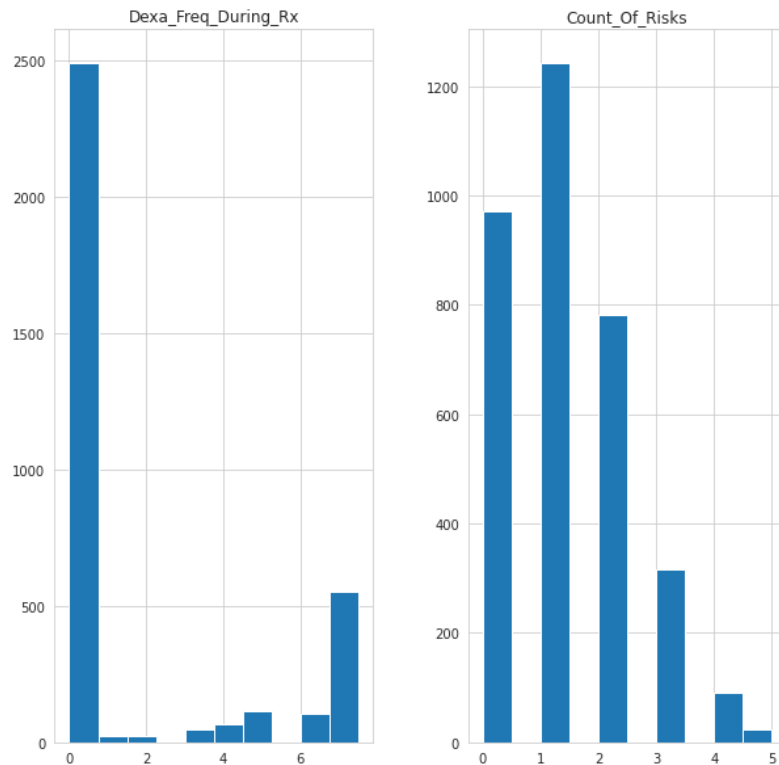
Assumptions:

- The patients were selected at random
- The variables were collected independently from each other



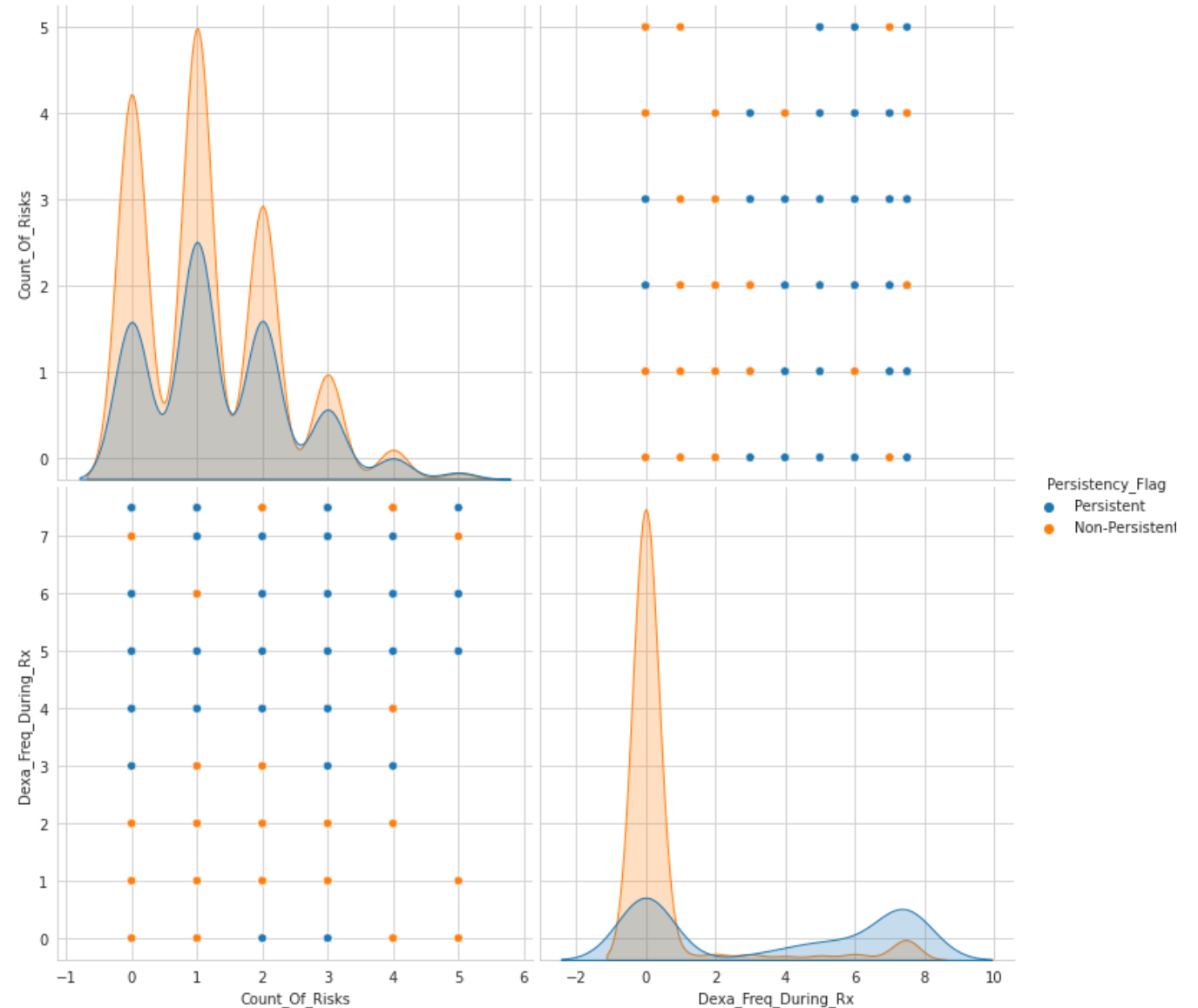
Quantitative Data Analysis

- There are only two quantitative data features, and they show no noteworthy features other than being both right-tailed skewed.
- Most important feature here is that most patients fall under 0-3 risk counts and less patients fall above a higher risk count.



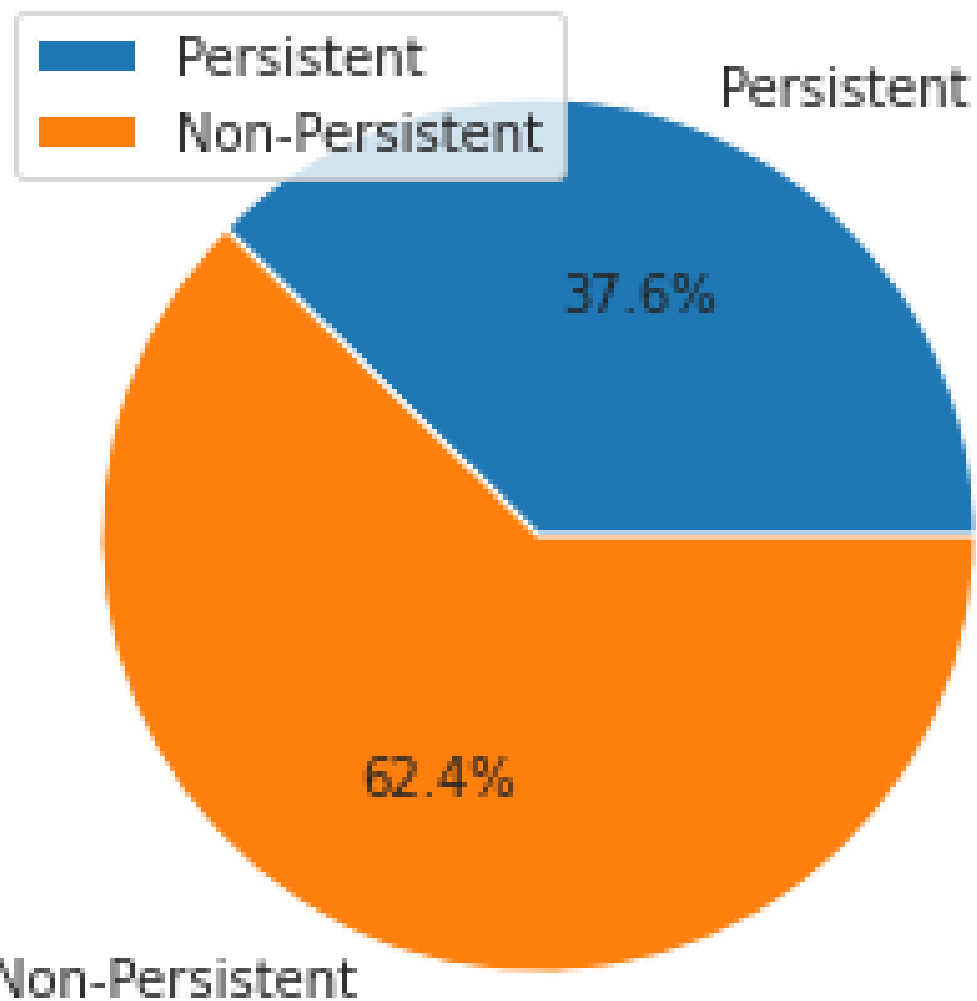
Quantitative Analysis

- There is considerable overlap between “Count_Of_Risks” and “Dexa_Freq_During_Rx” with “Persistency_Flag” which suggests a visual relation.
- The top graph suggests most patients (regardless of flag) has a lower risk count.
- Considerably more non-persistent patients had fewer Dexa scans during the medicating period than persistent patients.

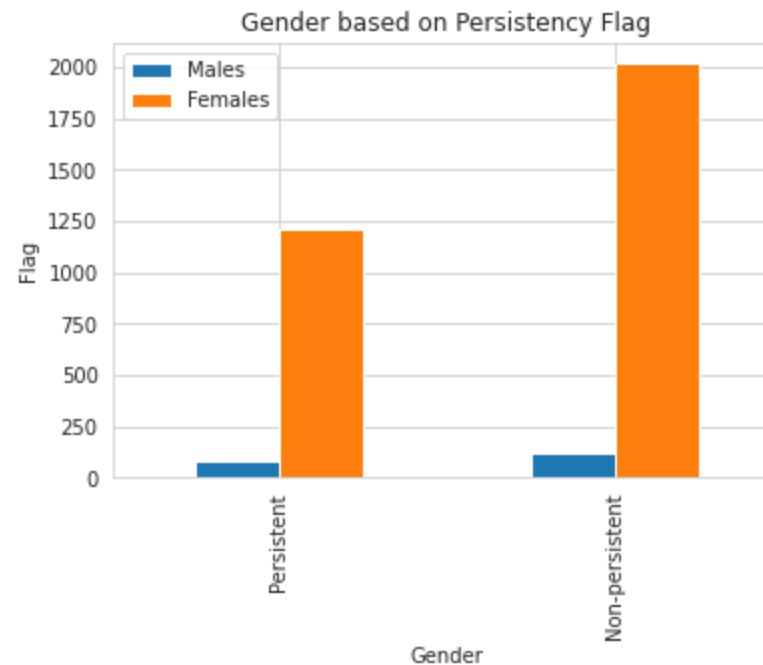
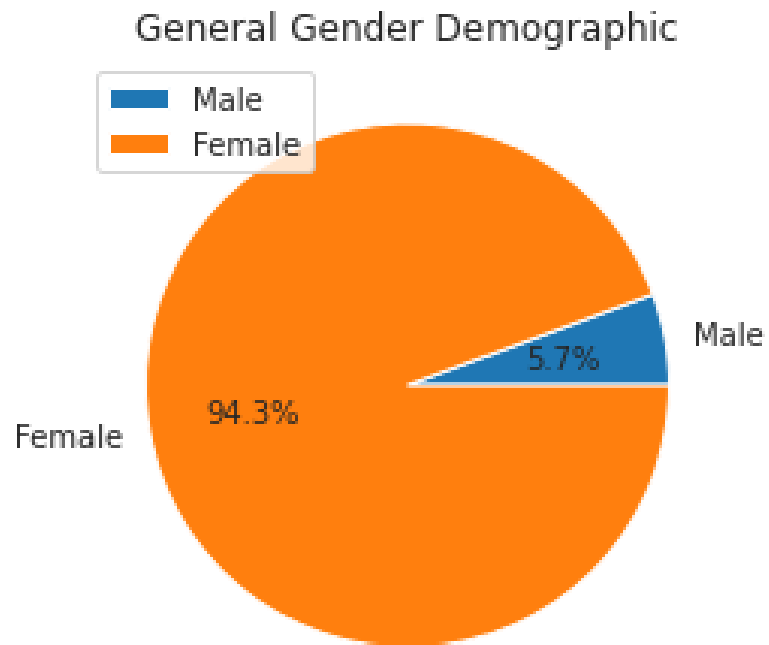


Patient Demographic- Patient Flag Persistency

Flag persistency of the overall demographic



Patient Demographic- Gender

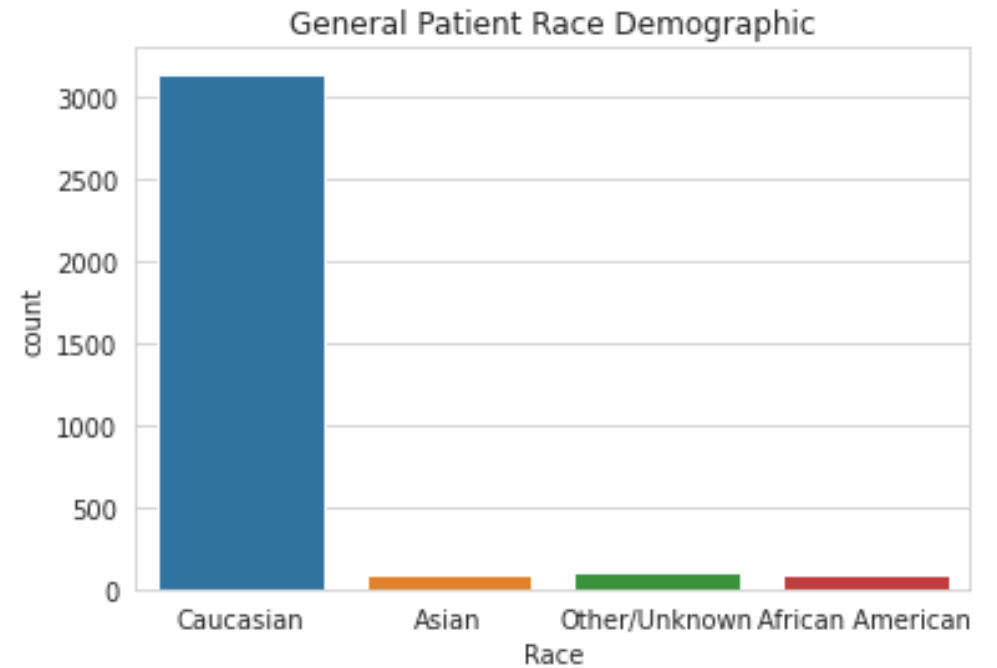
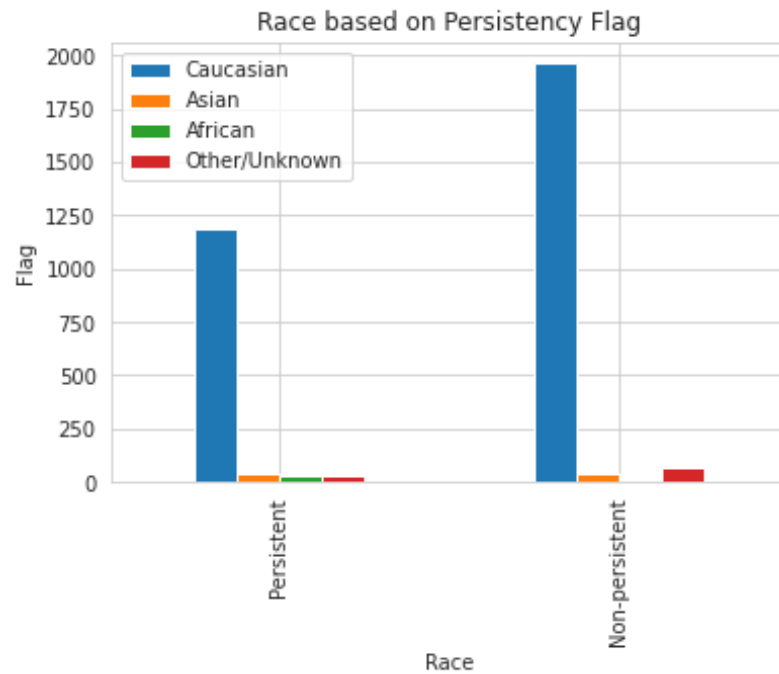


	Gender:	
Flag:	Male	Female
Persistent	77	1212
Not Persistent	117	2018

There are nearly 17x more females than males overall with both flag divisions having overall more females than male patients.

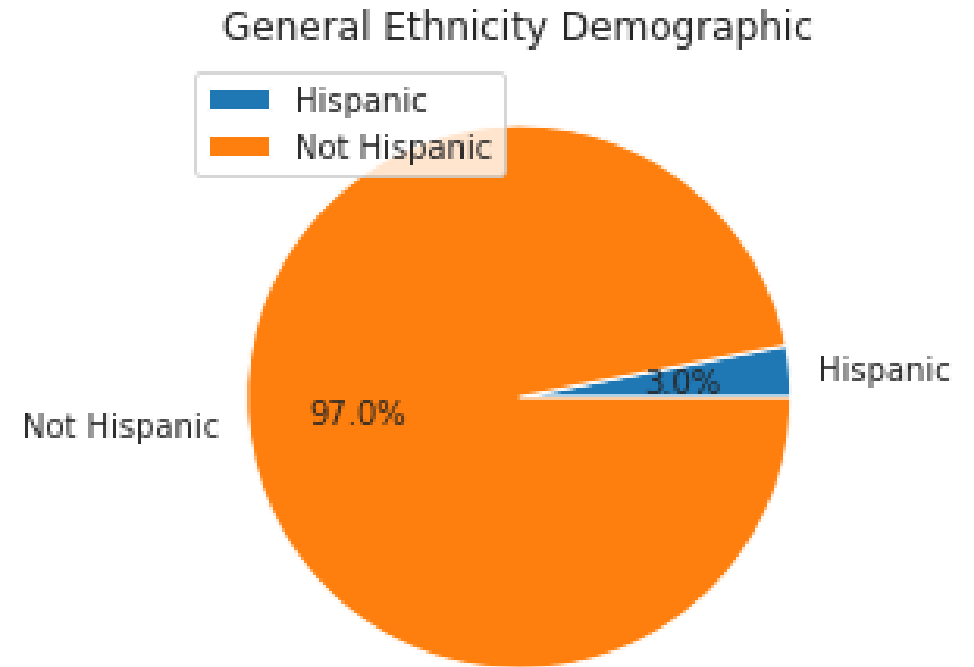
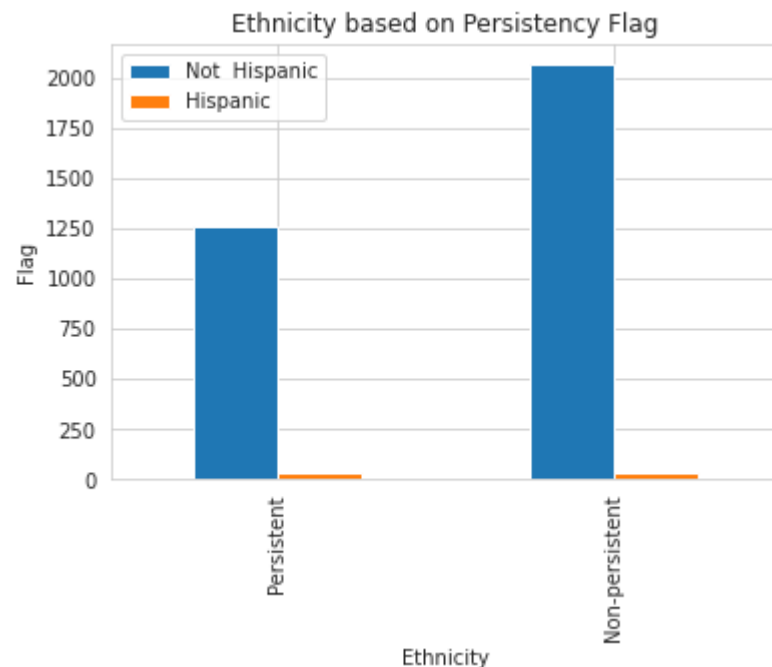
Patient Demographic- Ethnicity

- The vast majority of patients overall (and for each flag division) are belonging to a Caucasian ethnicity.

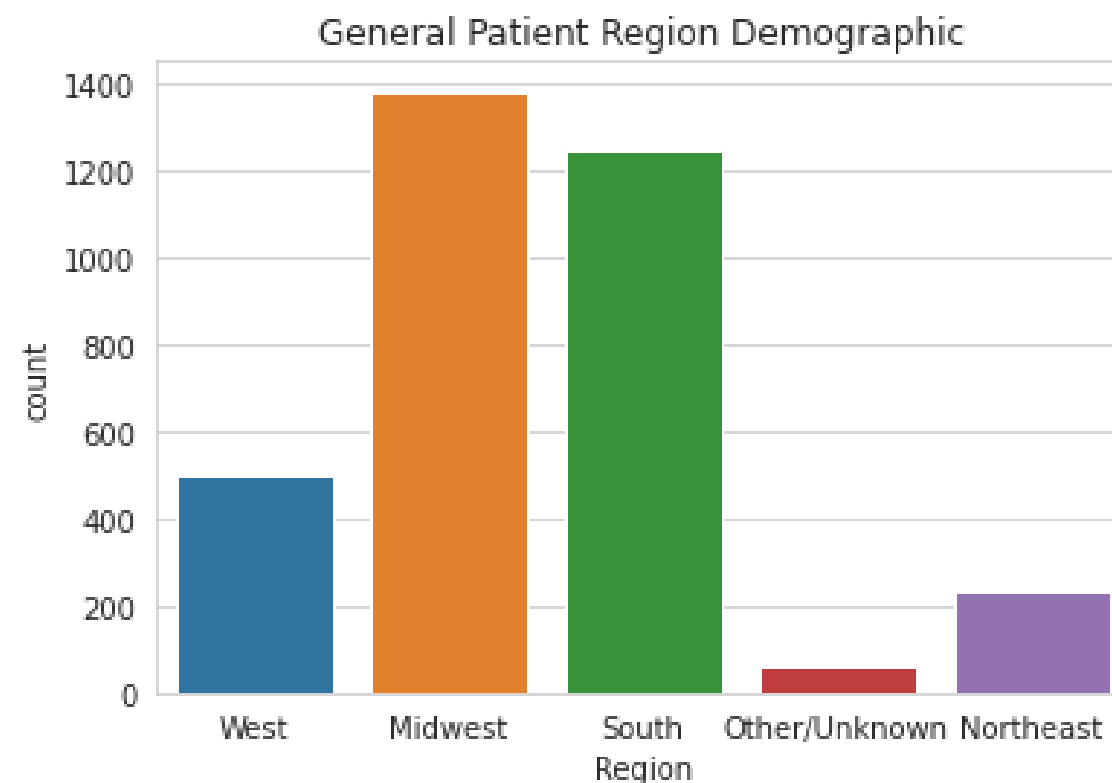
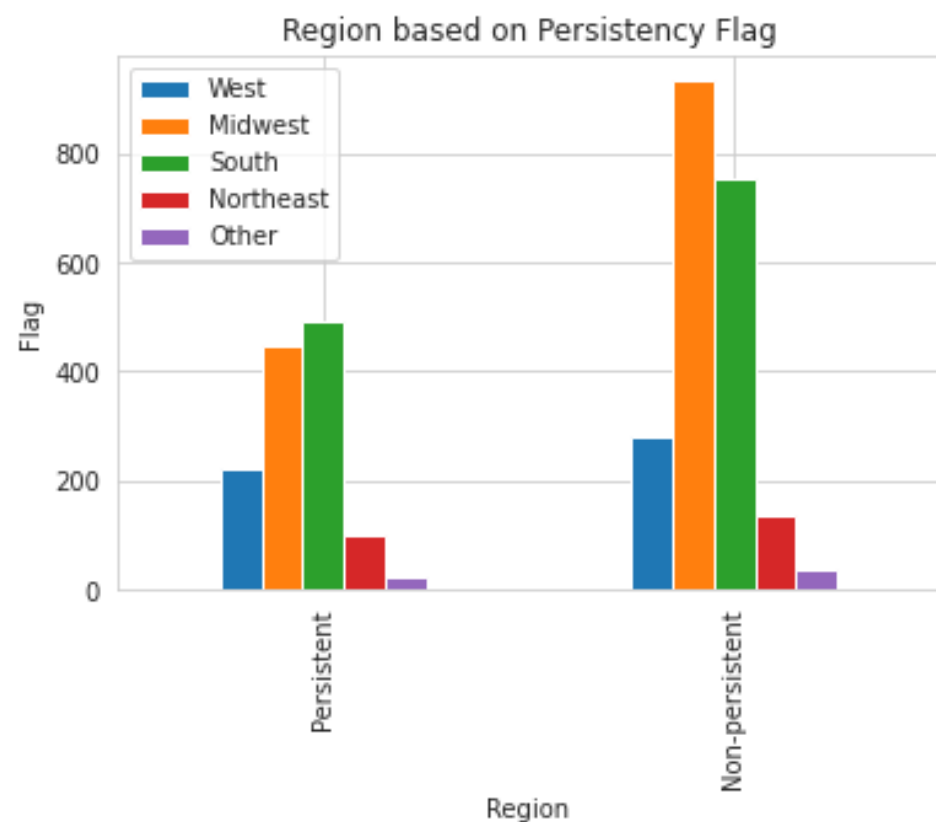


Patient Demographic-Ethnicity

- Overwhelmingly proportion of patients overall (and for both flag divisions) are not Hispanic.
- Specifically, non-Hispanics outnumber Hispanics by nearly 49x.

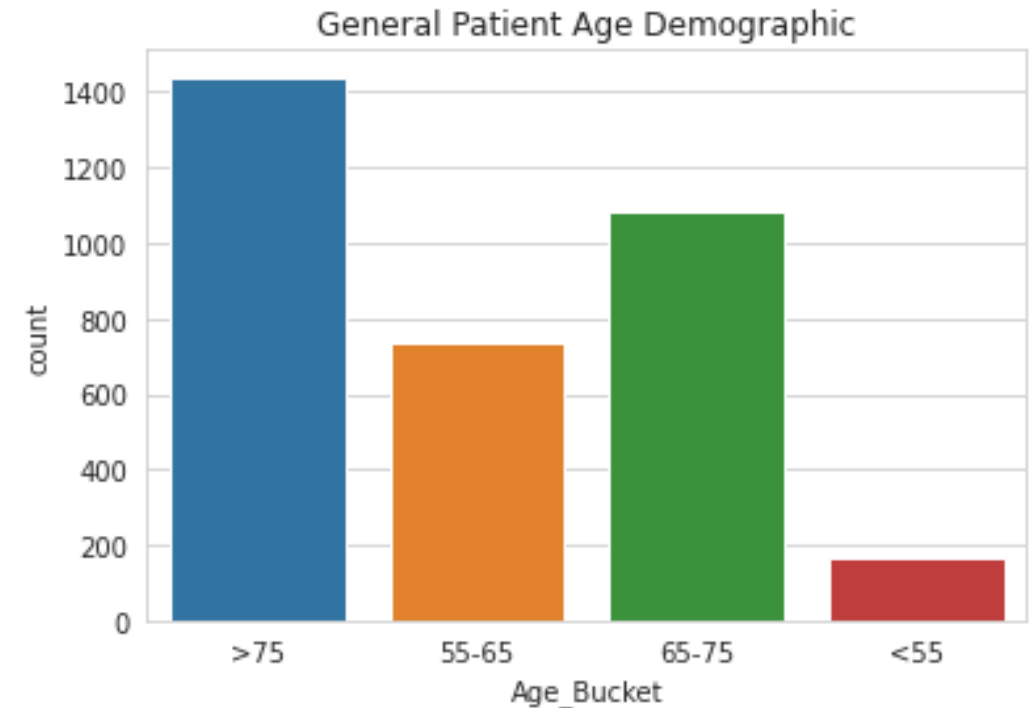
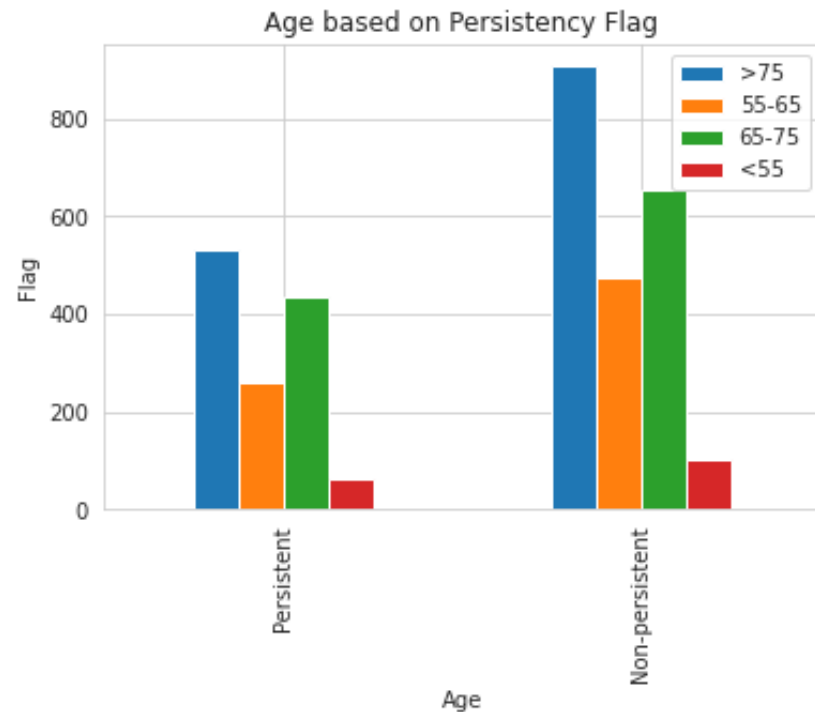


Patient Demographic- Region



Patient Demographic- Age

- Most patients are older with the majority falling older than 75 followed by a considerable number of patients between 65-75.
- There is not a considerable number of patients under 65.



Categorical Data Analysis

- A chi-square test is used to determine the association of categorical variables to flag persistency.
- Alpha value that is used is 0.05 (standard)
- The test calculates a p-value. If this value is ≤ 0.05 , we reject the null and believe the variables are associated with each other. If the p-value is > 0.05 , we fail to reject the null and believe the variables have no association with one another.
- H_0 (null): The two categorical variables are independent
- H_1 (alternative): The two categorical variables are dependent

Categorical Data Analysis

- Around 44 of the 69 features or 64% of the variables are said to be dependent with flag persistency (2 quantitative and 67 categorical variables).
- It will be these 44 features that will be going into our machine learning model.
- The picture to the right, listing some of the features, is not comprehensive

```
'Region', 'Ntm_Speciality', 'Ntm_Specialist_Flag',  
'Ntm_Speciality_Bucket', 'Gluco_Record_During_Rx', 'Dexa_During_Rx',  
'Frag_Frac_During_Rx', 'Change_T_Score', 'Change_Risk_Segment',  
'Adherent_Flag', 'Idn_Indicator', 'Injectable_Experience_During_Rx',  
'Comorb_Encounter_For_Screening_For_Malignant_Neoplasms',  
'Comorb_Encounter_For_Immunization',  
'Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx',  
'Comorb_Vitamin_D_Deficiency',  
'Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified',  
'Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx',  
'Comorb_Long_Term_Current_Drug_Therapy', 'Comorb_Dorsalgia',  
'Comorb_Personal_History_Of_Other_Diseases_And_Conditions',  
'Comorb_Other_Disorders_Of_Bone_Density_And_Structure',  
'Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias',  
'Comorb_Osteoporosis_without_current_pathological_fracture',  
'Comorb_Personal_history_of_malignant_neoplasm',  
'Comorb_Gastro_esophageal_reflux_disease',  
'Concom_Cholesterol_And_Triglyceride_Regulating_Preparations',  
'Concom_Narcotics', 'Concom_Systemic_Corticosteroids_Plain',  
'Concom_Anti_Depressants_And_Mood_Stabilisers',  
'Concom_Fluoroquinolones', 'Concom_Cephalosporins',
```



Recommendations (Machine Learning Models)

- We will be using a binary classification algorithms to predict flag persistency.
- Binary classification is used for data where there are only two outcomes which takes on either a “0” or a “1.” in our case, the cases match non-persistency and persistency, respectively.
- Visually represented by the discrete Bernoulli distribution.

Models to consider:

- Logistic Regression
- Support Vector Machines
- Simply Bayes
- Decision Trees



Machine Learning Results


- A multiple linear regression model was selected due to a lack of comparable test datasets needed for other choices i.e., decision trees. Before we dive into the results, let us state the assumptions:
 1. Linearity
 2. Independence
 3. Homoscedasticity
 4. Normality



Machine Learning Results

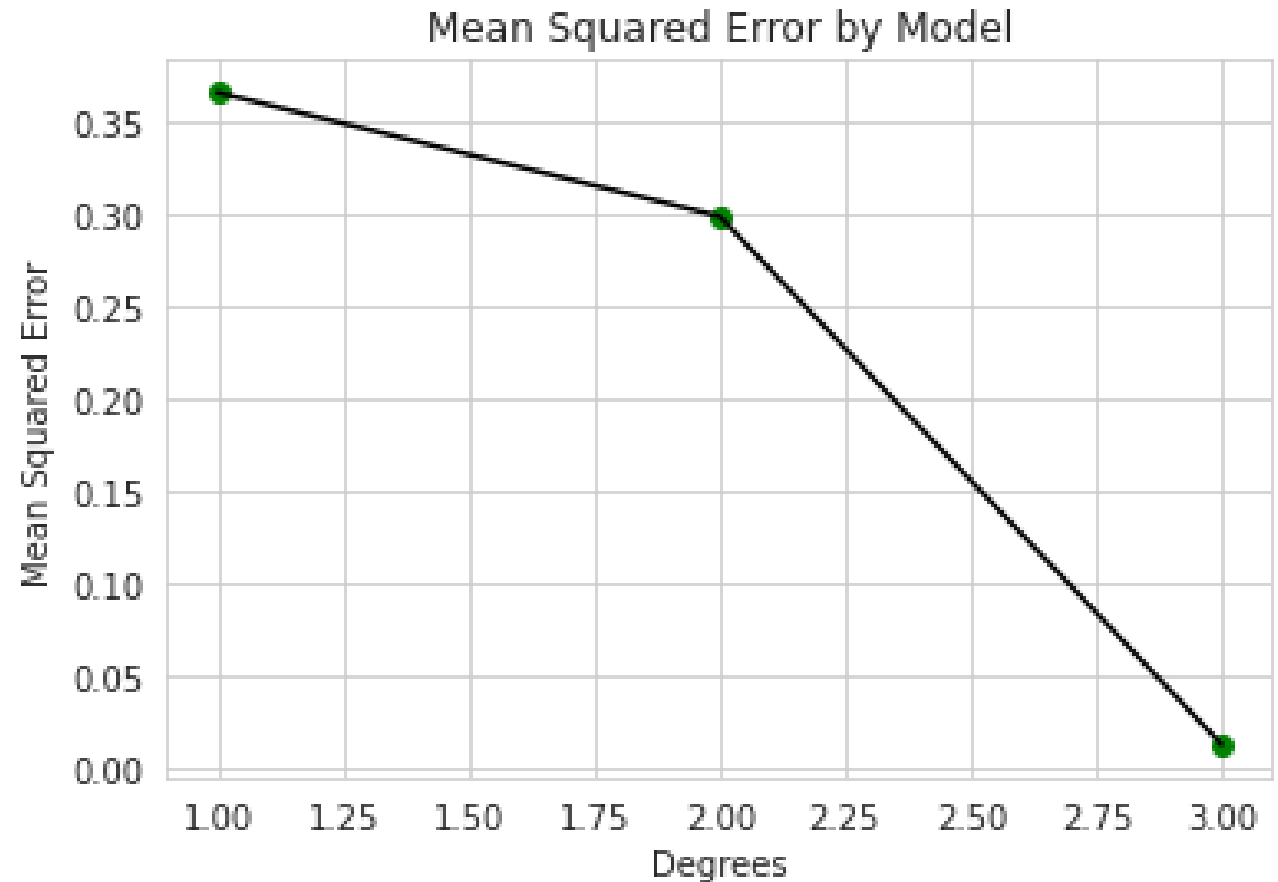
- The multiple linear regression model has proved to not be a good fit. Here is why:
 1. The coefficient of determination is about 0.4 (weak). For comparison, a 1.0 represents a perfect fit. This might be an indication of underfitting and that we need to use a more complicated model.
 2. Some of the individual R^2 's for the independent variables turned out negative.
 3. The assumptions might not be applicable with our data.

Solution: Fit the data with a multivariate polynomial regression model with a degree that is > 1 .



Machine Learning Results

- The new model created using a multivariate polynomial model at degree three shows the most promising fit for the data, producing a mean squared error of about 0.012 and a root mean squared error of roughly 0.11 which is considerably better than the linear model at about 0.14 and 0.39, respectively.
- A degree three polynomial was chosen to limit overfitting from higher-order terms.
- As a standard, root mean squared error is usually a better indicator of model fit than the alternative.





Limitations & Expansion

If I were to repeat this project, this is what I would have done differently:

- Conducted my quantitative data analysis after transforming my dataset (consisting of mostly categorical data) into numerical values.
- Tested out assumptions before jumping into a model.
- Implemented a logarithmic model and compared the results with the ones obtained from the multivariate polynomial regression model.

It would be the most optimal to have multiple datasets like ours to test for drug persistency. If we had this, we could:

- Implement more sophisticated classification algorithms.
- Implement boosting models i.e., discrete AdaBoost would be a strong choice as it deals with binary classification problems.



Sources

- <https://enjoymachinelearning.com/blog/multivariate-polynomial-regression-python/>
- <https://data36.com/polynomial-regression-python-scikit-learn/>
- <https://www.linkedin.com/pulse/my-first-exploratory-data-analysis-project-dr-ragini-selukar/>
- <https://stephenallwright.com/rmse-vs-mse/#:~:text=RMSE%20is%20one%20of%20the,is%20often%20preferred%20over%20MSE.>
- https://www.linkedin.com/posts/emilyyao1_introductory-data-cleaning-activity-7013640766505820160-0MWM?utm_source=share&utm_medium=member_desktop
- <https://machinelearningmastery.com/adaboost-ensemble-in-python/>
- <https://intellipaat.com/blog/what-is-linear-regression/#Multiple-Linear-Regression-Model>
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- <https://realpython.com/linear-regression-in-python/#polynomial-regression>



Data Glacier

Your Deep Learning Partner

Thank You

The End.