Exploratory Data Analysis

Data Science Persistency of a Drug Project

Data Force

9/26/22

Agenda

- Executive Summary
- Approach
- EDA
- EDA Summary
- Proposed Model
- Final Model Build
- Conclusion

Team Information

Group Name: Data Force

Name	Email	Company	Country	Specialization
Anshi Mathur	anshimathur0325@gmail.com	Data Glacier	United States of America	Data Science
Lujain Saad	lujainsaadcs@gmail.com	Data Glacier	Saudi Arabia	Data Science
Prince Kumar Lat	princek.iitk@gmail.com	Data Glacier	Canada	Data Science
Mohamed Amine Kina	kinaamine@gmail.com	Data Glacier	Germany	Data Science

Background/Information

In medical terms, the persistency of a drug refers to the extent while the patient acts accordingly to the effects of a drug. Patients lose persistency when the drug no longer has an affect on them.

ABC Pharma is a pharmaceutical company who aims to automate identification of the persistency of a drug.

In this data analysis, the various factors of a patient are evaluated to determine if they have an affect on the persistency.

At the end of this project, we will suggest a model to follow for deployment.

Approach

Given: 1 dataset with information about each patient and whether they had persistency with a drug.

Method of Approach:

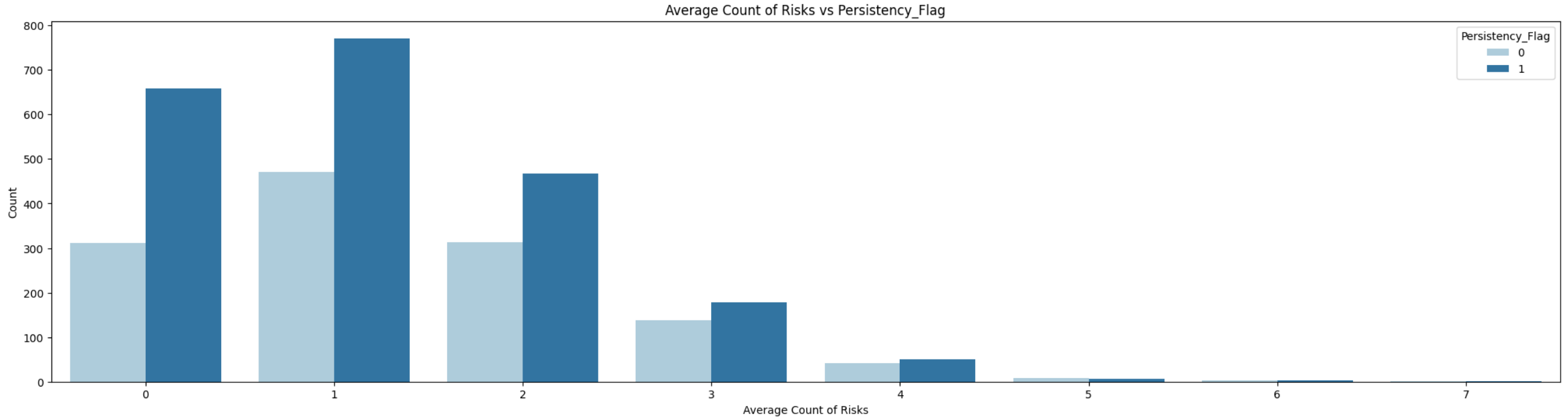
1. Look for null or missing values.
2. Identify values that are improbable.
3. Create visualizations of the data.

Assumptions:

1. Each patient adheres to a specific ID and do not have multiple.
2. Data was selected at random from a population and not targeted specifically based on a patient's previous persistency.

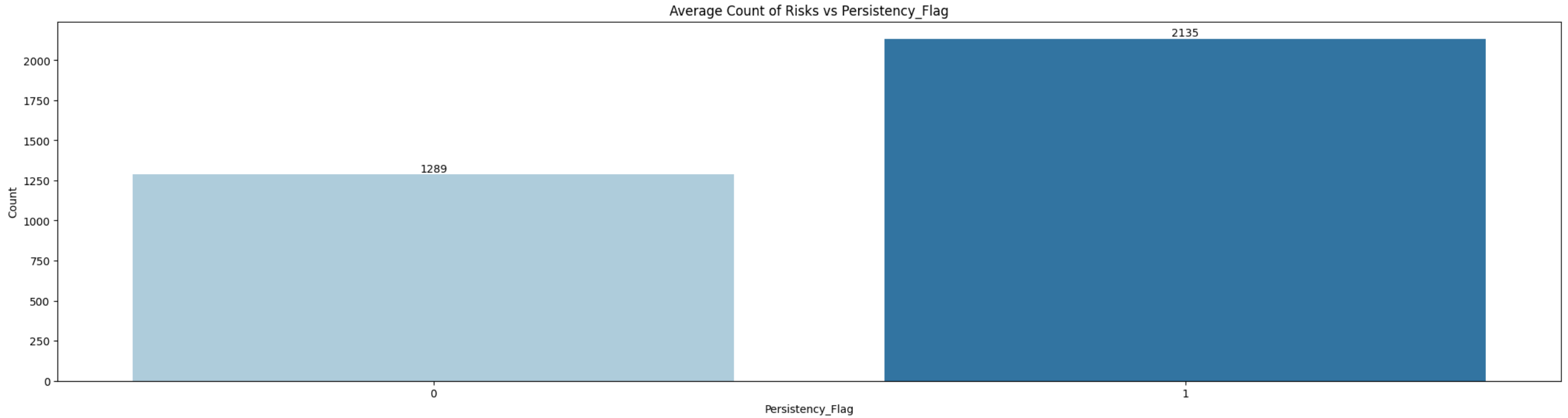
EDA

Risks vs. Persistency Flag



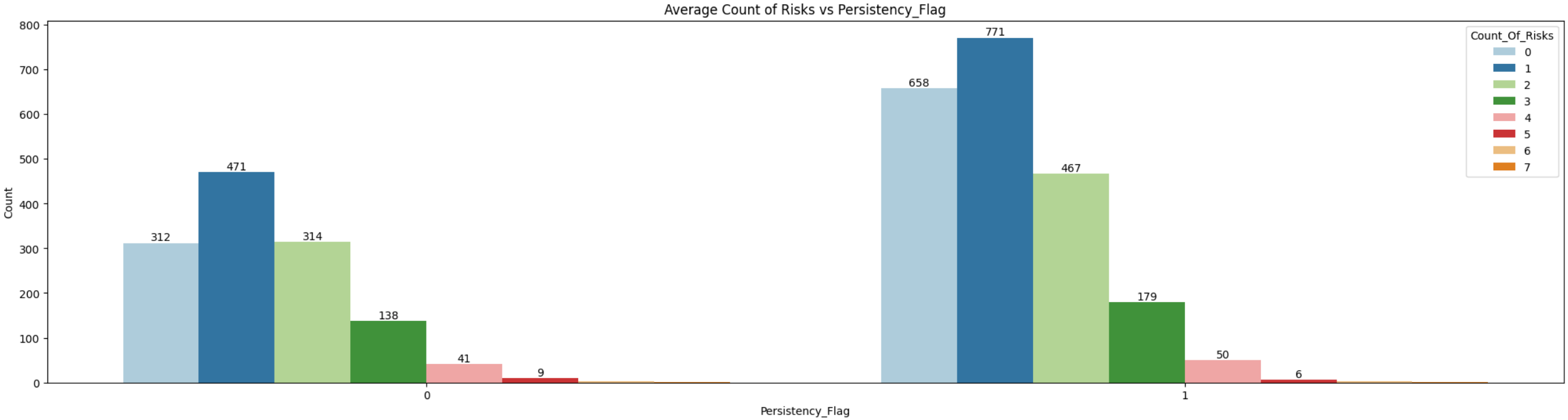
Using this graph, it is clear to see that more people were persistent when they had 0-1 risks. The greatest amount of non-persistent patients were at 1 risk, and as the risks increased, the gap between the two groups decreased.

Risks vs. Persistency Flag



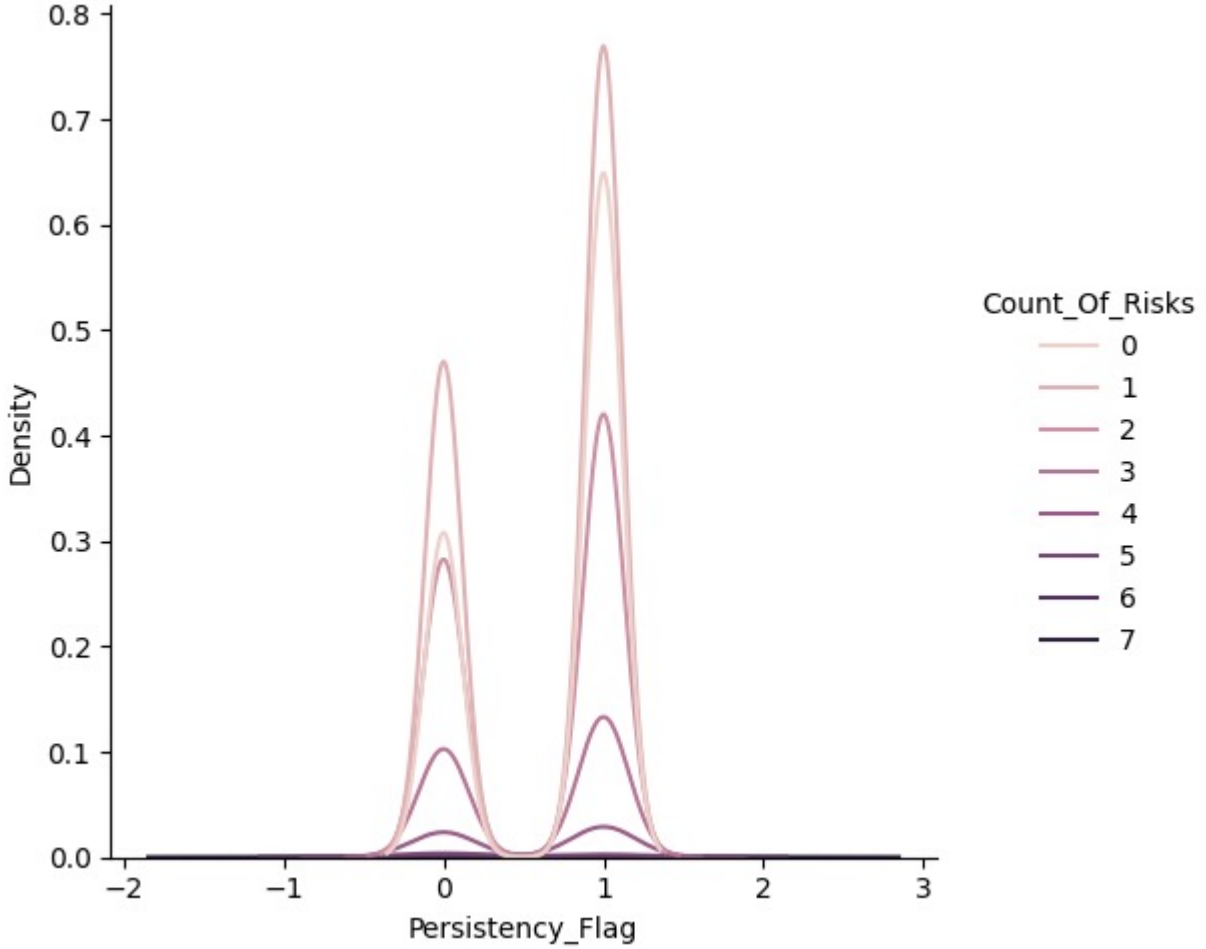
With this data table, it is clear that the data collected has more information about the people that were persistent than the people that were not. Therefore, the data itself may seem to favor the persistent group of people but in reality that is due to the collection of data.

Risks vs. Persistency Flag



Overall, both the non-persistent and the persistent group of people follow a relatively similar distribution. Overall, in the non-persistent group side, more people have 2 risks than no risks, which indicates that risks of diseases may negatively impact the persistency of a drug. For the persistent group, both 0 and 1 risks dominate over the other counts of risks.

Risks vs. Persistency Flag

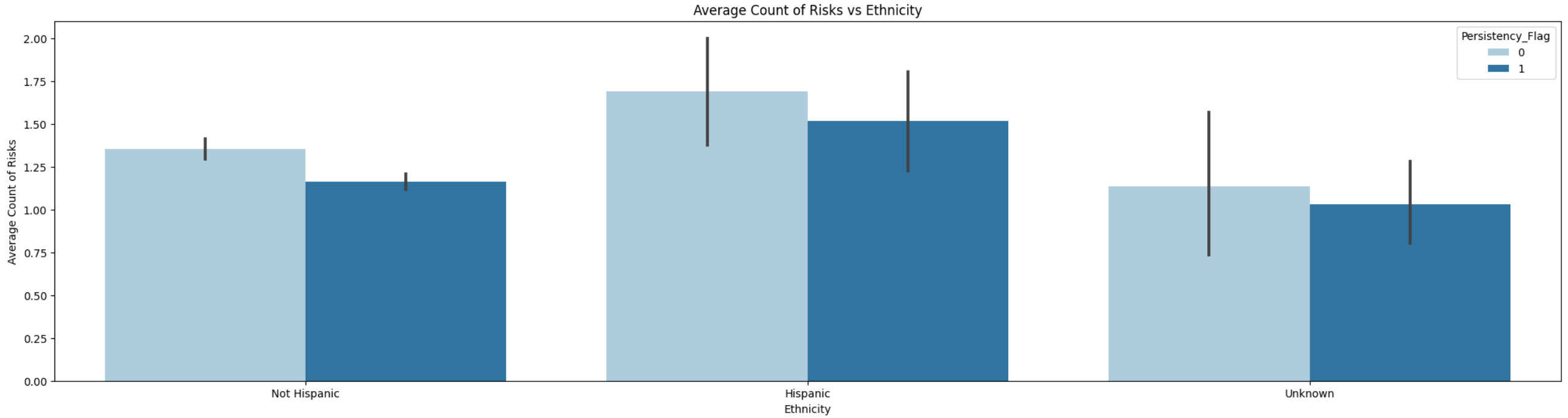


Using this data, it is clear to see that there is a greater amount of people with no risks who have experienced persistency compared to the amount who have not.

This suggests that in order for a drug to be persistent on a patient, the patient must have little to no risks.

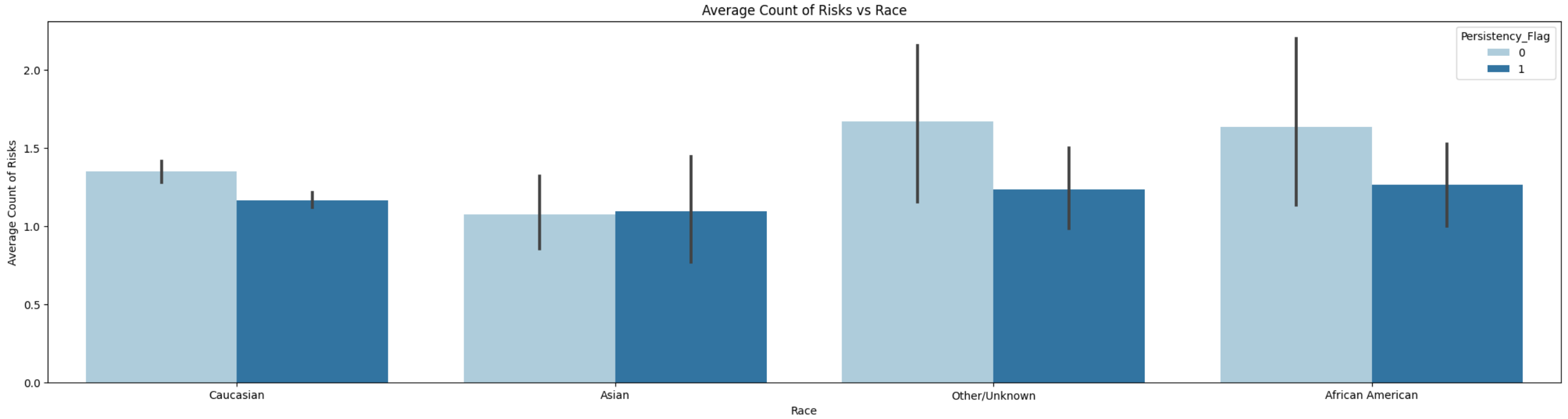
Overall, risks pose a significant impact on the persistency of a drug on a patient.

Risks vs. Ethnicity



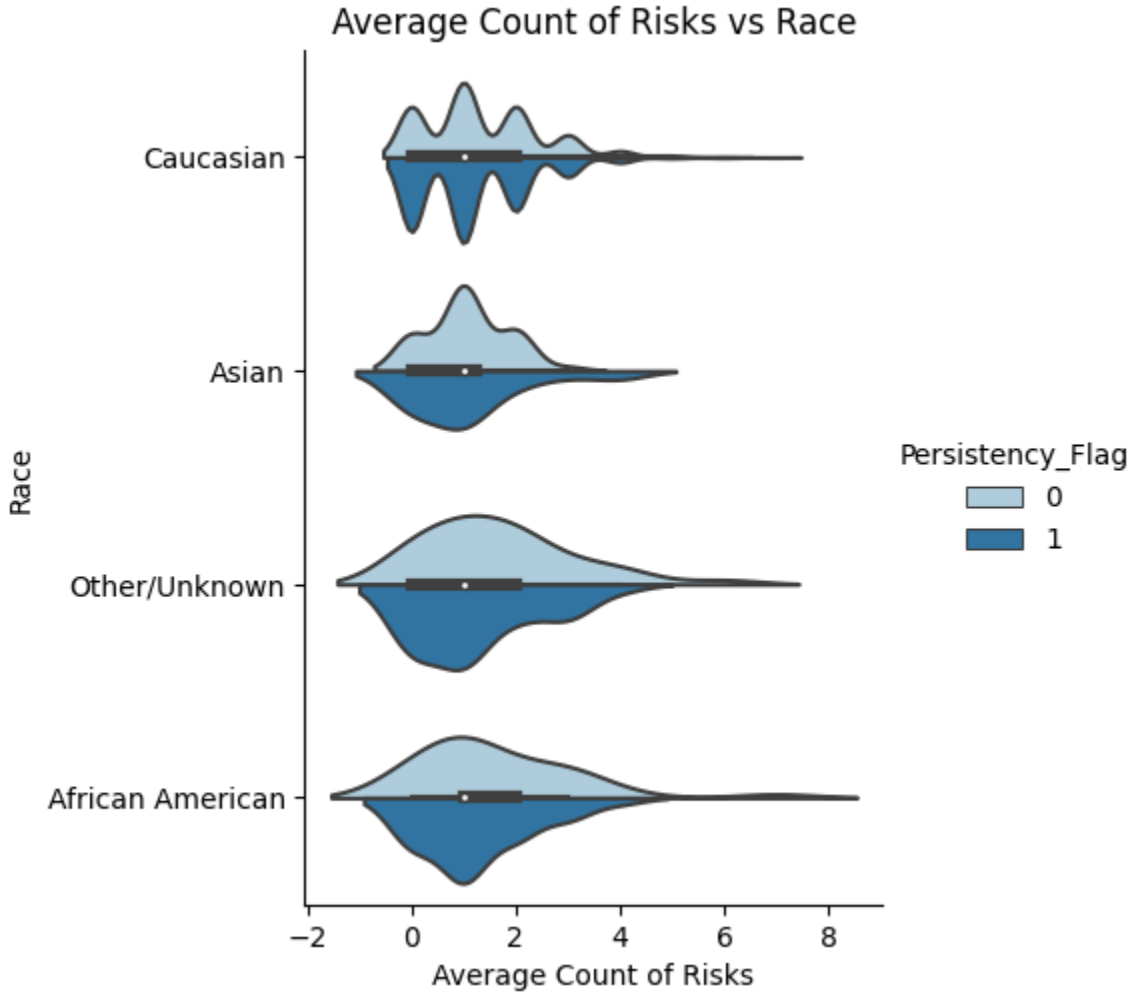
Overall, on average the Hispanic community faces a greater number of risks and a greater chance of not being persistent to a drug. In the graph above, the Hispanic group who does not experience persistency is the greatest column, indicating that the Hispanic community may find it difficult to increase their persistency with drugs.

Risks vs. Race



Overall, the Asian race seems to have a strong balance between people who have and have not experienced persistency. For the other races, however, people who don't have persistency with drugs dominate. This is especially emphasized for the Other category and the African American category. On average, the African American community has a greater number of risks, which leads them to face a greater challenge in finding the persistency of a drug.

Risks vs. Race

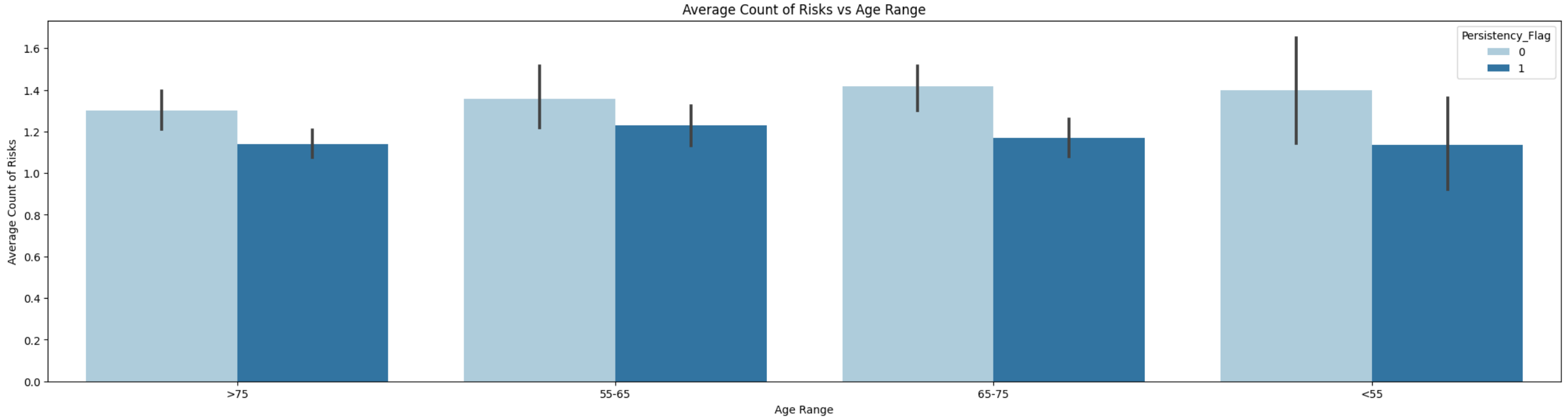


This chart on the left is another graphical representation of the average count of risks and the race of a patient.

It is clear that on average, the Caucasian race faces more variability for persistency within their group.

For the African American group, as the risks increase, the non-persistent group grows slightly larger than the persistent group, suggesting that they face less persistency as a whole.

Risks vs. Age



The groups of ages <55 and 65-75 seem to have a greater disparity between persistent and non-persistent patients. Group 55-65 has a reduced distance between the two groups. Using this data, it is clear that younger patients may not have persistency towards a drug and older people as well.

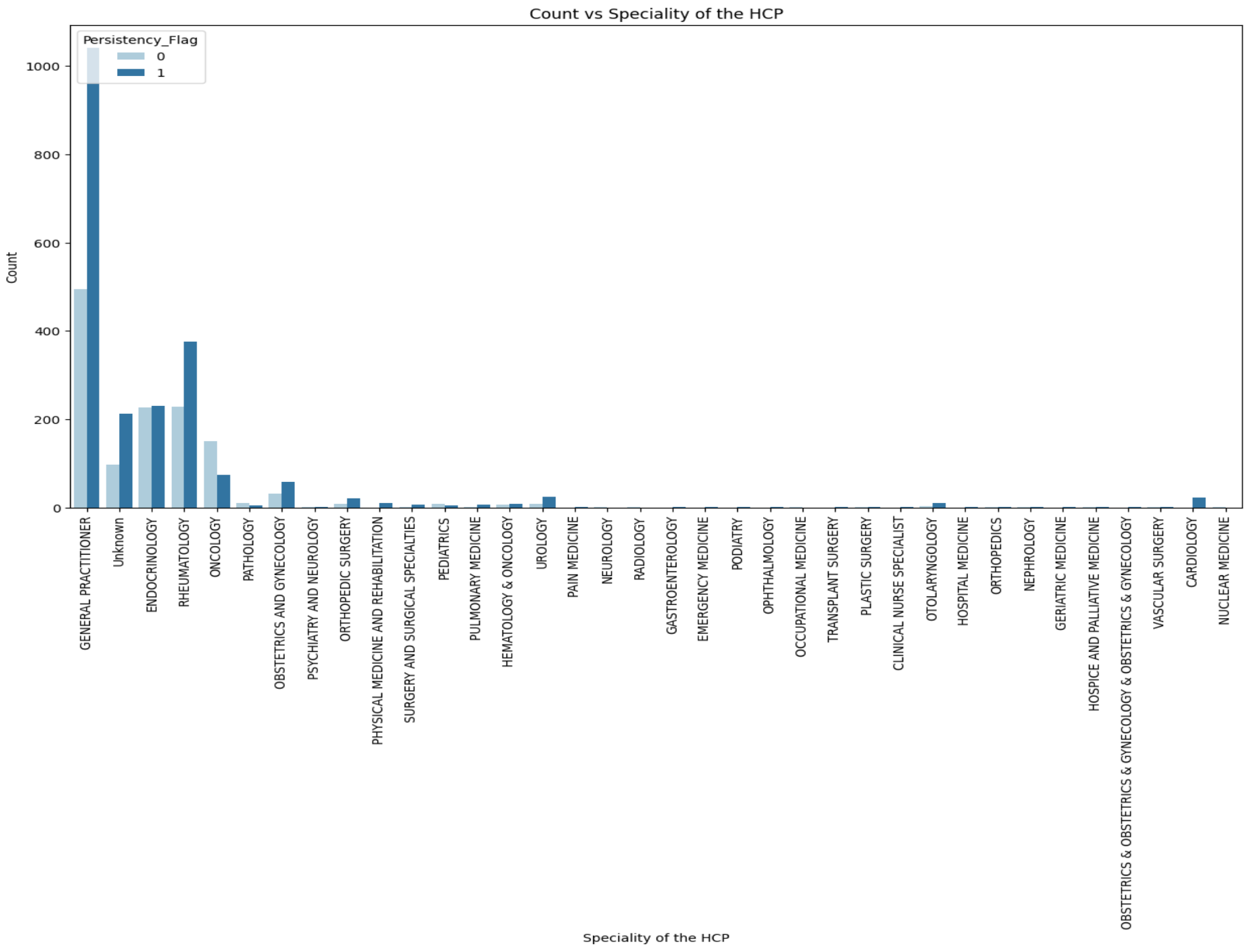
Specialty vs Count

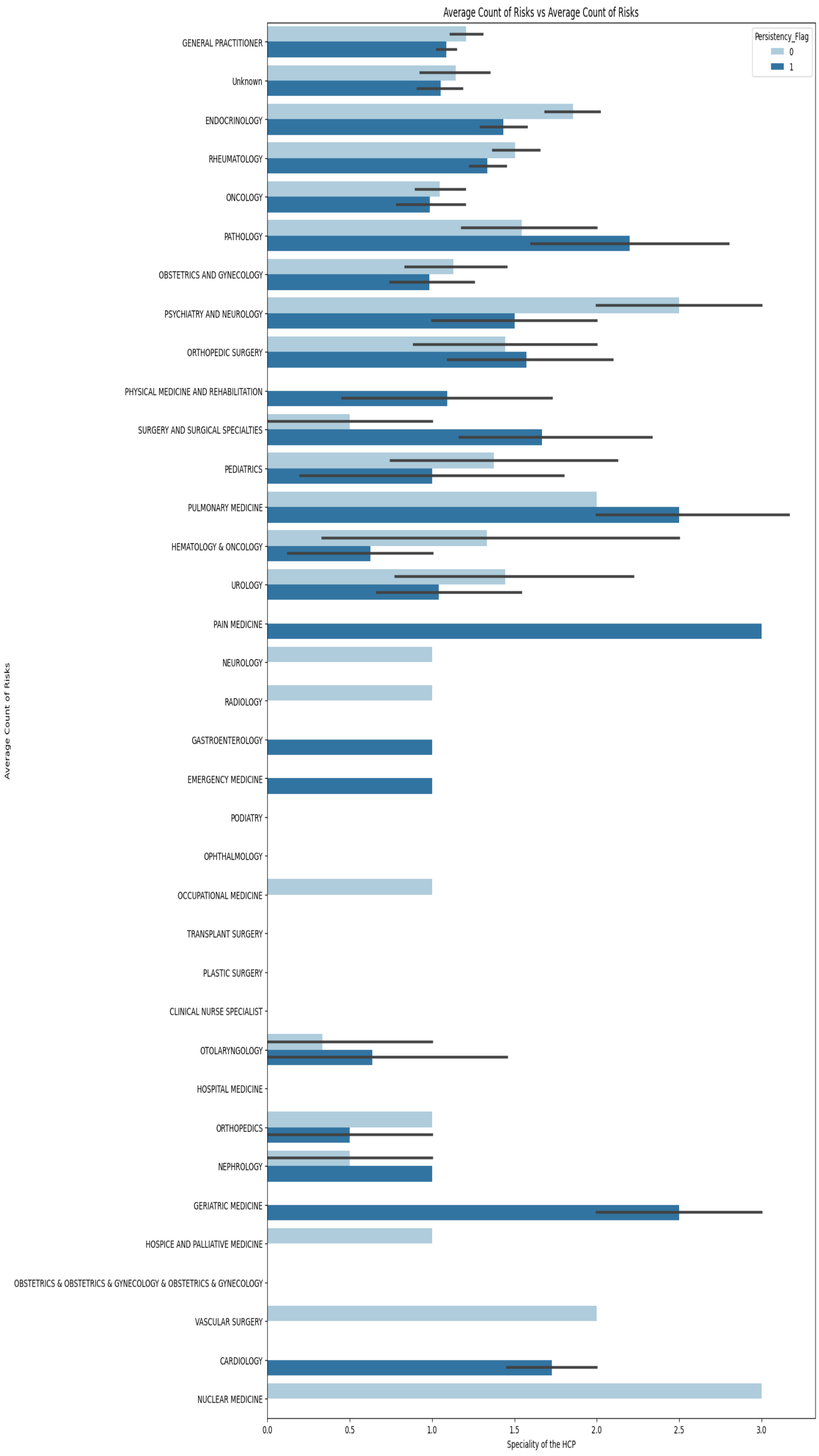
The data from this graph represents the specialty of the HCP that prescribed the drug.

Clearly, the general practitioner prescribed the most medicines.

The endocrinologist has about 49% of their drugs not be persistent.

Oncology also has a minimal number of drugs being persistent. This could also be due to cancer not having a specific cure.




Average Count of Risks vs Average Count of Risks
Persistency_Flag
0
1
GENERAL PRACTITIONER
Unknown
ENDOCRINOLOGY
RHEUMATOLOGY
ONCOLOGY
PATHOLOGY
OBSTETRICS AND GYNECOLOGY
PSYCHIATRY AND NEUROLOGY
ORTHOPEDIC SURGERY
PHYSICAL MEDICINE AND REHABILITATION
SURGERY AND SURGICAL SPECIALTIES
PEDIATRICS
PULMONARY MEDICINE
HEMATOLOGY & ONCOLOGY
UROLOGY
PAIN MEDICINE
NEUROLOGY
RADIOLOGY
GASTROENTEROLOGY
EMERGENCY MEDICINE
PODIATRY
OPHTHALMOLOGY
OCCUPATIONAL MEDICINE
TRANSPLANT SURGERY
PLASTIC SURGERY
CLINICAL NURSE SPECIALIST
OTOLARYNGOLOGY
HOSPITAL MEDICINE
ORTHOPEDICS
NEPHROLOGY
GERIATRIC MEDICINE
HOSPICE AND PALLIATIVE MEDICINE
OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY
VASCULAR SURGERY
CARDIOLOGY
NUCLEAR MEDICINE
Average Count of Risks
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Speciality of the HCP

Specialty vs Count

The previous slide uses another graphical representation of a HCP's specialty and the average count of risks.

Some of these specialties are rare to identify, which may affect the data analysis. With more data on less popular specialties, we may be able to draw conclusions based on the HCP's specialty.

Overall, the specialty of a person's HCP has the potential to have an affect on the persistency of a drug for a patient.

EDA Summary

Out of the many factors that may affect a patient's persistency to a drug, having risks, being of a certain race, ethnicity, age group, and the specialty of the HCP have the strongest affect on a patient's persistency towards a drug.

Now that the factors have been identified, we will be able to use a model to create an automated process of which can predict if a drug would be persistent on a patient or not.

For a technical user:

Out of the many models of machine learning algorithms, the random forest classifier model would be the best one to use in this scenario. With additional testing, the random forest classifier model was the most effective and accurate for predicting whether or not a patient would have persistency with a drug.

Model Building

Proposed Model

In order to create the most effective machine learning program, it is important to use the most effective model.

We tested the data with Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and XGBoost and tested their accuracy in predicted the right score.
Accuracy rates-
Logistic Regression: 0.810
Random Forest Classifier:0.815
Decision Tree Classifier:0.745
XGBoost: 0.809

It is clear that the Random Forest Classifier model is the most effective.

Final Model Build

To build the most effective model, we first used a cleaned dataset.

We made sure that the columns of the dataset were integer values when they could be, so that the machine learning module can be the most accurate without the data staying as strings.

To help with columns containing multiple string values, we created dummies to split the data apart, once again to help the algorithm learn smoother.

We then created a new dataset without the correlation column.

After all of this, we were able to apply the Random Forest Classifier algorithm and create a .pkl file out of it.

View of Model.py

The picture on the right represents the model we built using Python.

This model creates a .pkl file that can be used when making an application to help predict the persistency of a drug.

model.py > ...

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import pickle
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier

#Drop null values.
df_healthcare = pd.read_csv("Healthcare_Dataset.csv", sep=",")
df_healthcare.dropna()
#Replace string variables with numbers to make the module more accurate.

for i in df_healthcare.columns:
    if(df_healthcare[i].unique().shape[0] ==2):
        df_healthcare[i] = df_healthcare[i].map({df_healthcare[i].unique()[0]:0, df_healthcare[i].unique()[1]:1})
df_healthcare_final = pd.get_dummies(df_healthcare, columns=['Race','Ethnicity','Region','Age_Bucket','Ntm_Speciality','Ntm_Speciality_Bu

df_corr = df_healthcare_final.corr()
corr_col = []
for i in range(len(df_corr.index)):
    for j in range(i):
        if((df_corr.iloc[i,j])>0.75):
            corr_col.append(df_corr.columns[j])
df_healthcare_final2 = df_healthcare_final.drop(corr_col, axis=1)

y_cut = df_healthcare_final2['Persistency_Flag']
x_cut = df_healthcare_final2.drop(['Ptid','Persistency_Flag'], axis=1)

trainxcut, testxcut, trainycut, testycut = train_test_split(x_cut,y_cut, test_size=0.3, random_state=19)

from sklearn.linear_model import LogisticRegression
rand = RandomForestClassifier()

rand.fit(trainxcut, trainycut)

pickle.dump(rand,open('model.pkl','wb'))
model=pickle.load(open('model.pkl','rb'))
```

Conclusion

Throughout this project, we were able to identify specific variables that affect the persistency of a drug, thorough the use of exploratory data analysis.

By using modules from Python including Seaborn and Matplotlib, we were able to create data visualizations to create a presentation to showcase our findings about the data.

After the EDA, we were able to try out various different machine learning algorithms and identified that Random Forest Classifier was the most effective one of the group.

Using this information, we were able to create a machine learning model, which creates a .pkl file that can be used in making an application for the persistency of a drug.

Thank you to Data Glacier for providing the resources for this project.

Thank You