Week 11: Data Science Healthcare Project

Name: Serhat Uğur

E-mail: ugur.serhat@outlook.com

Country: Turkey

University: Anadolu University

Specialization: Data Science

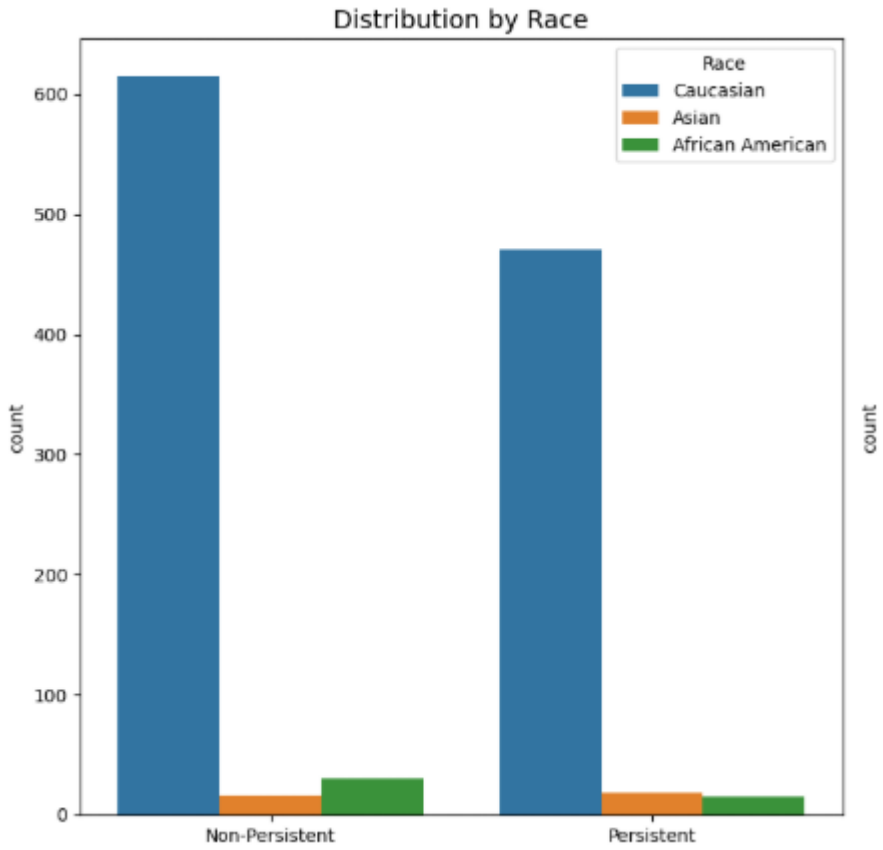
Problem Description

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 the identification process.

GitHub Repo Link: https://github.com/serhatugur/data-science-internship

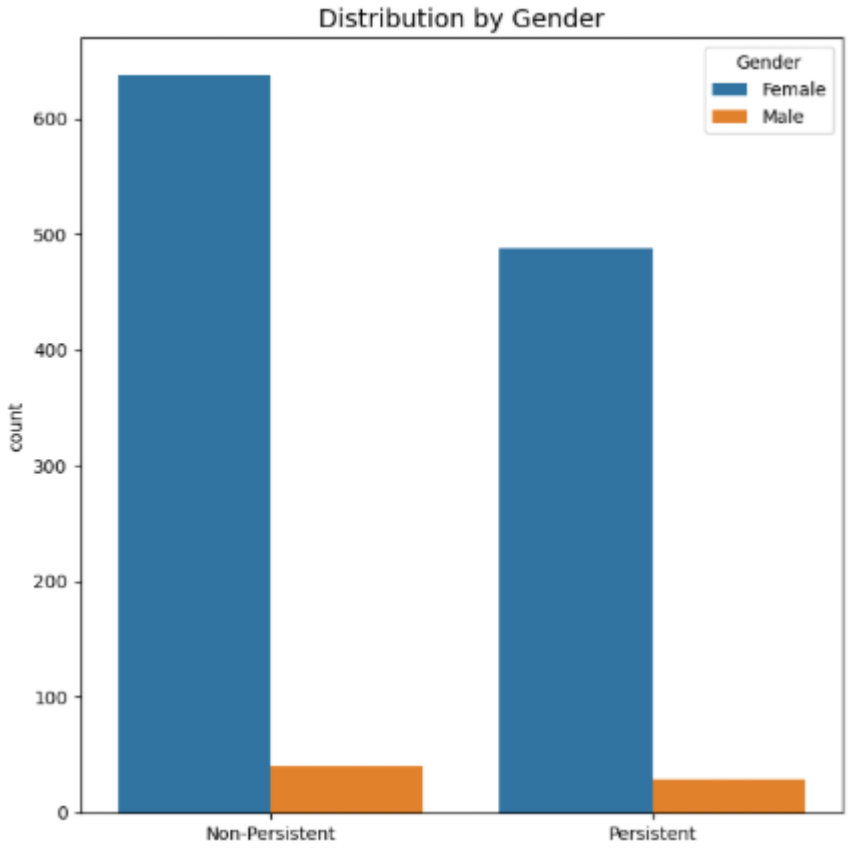
Exploratory Data Analysis

1) Distribution by Race



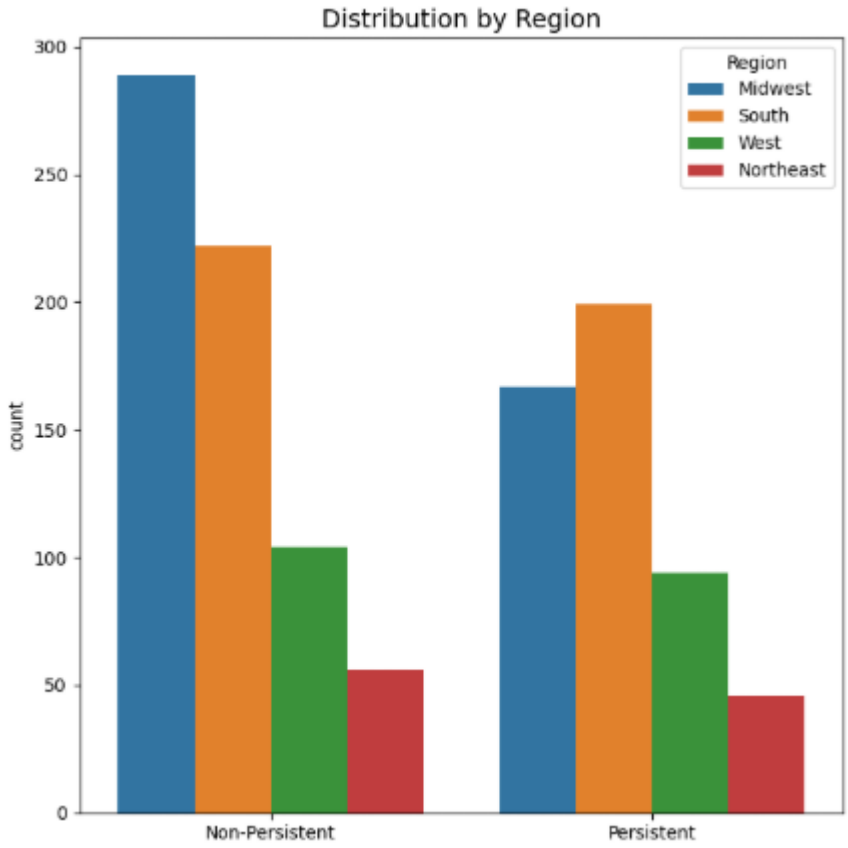
Caucasian is the most common race in the study.

2) Distribution by Gender



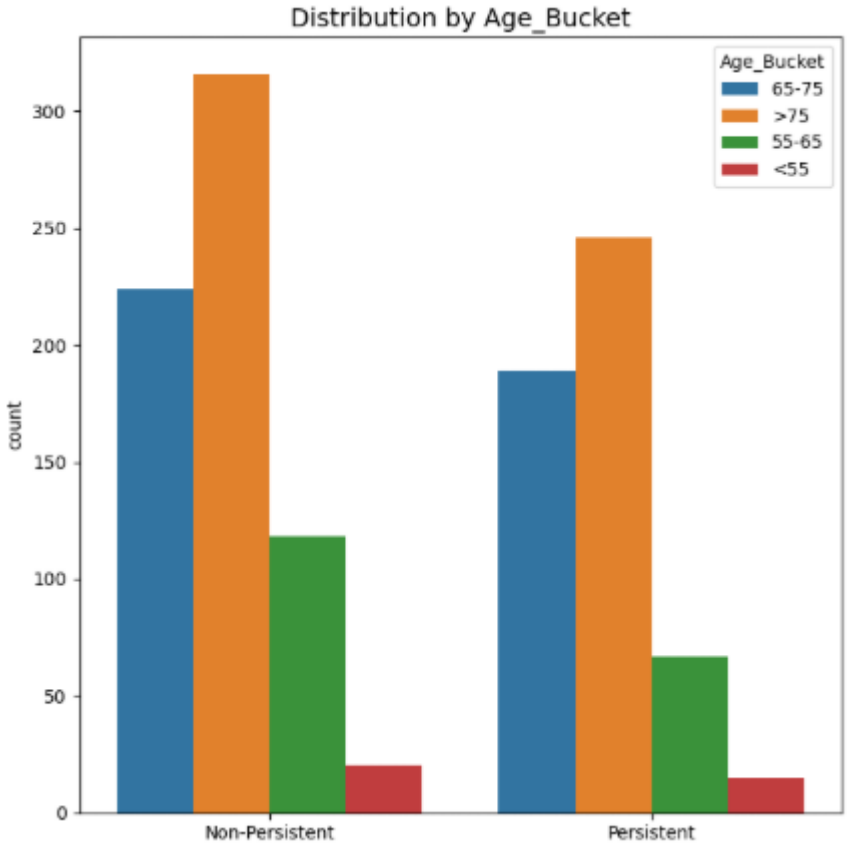
In both groups, females number much more than the number of men.

3) Distribution by Region



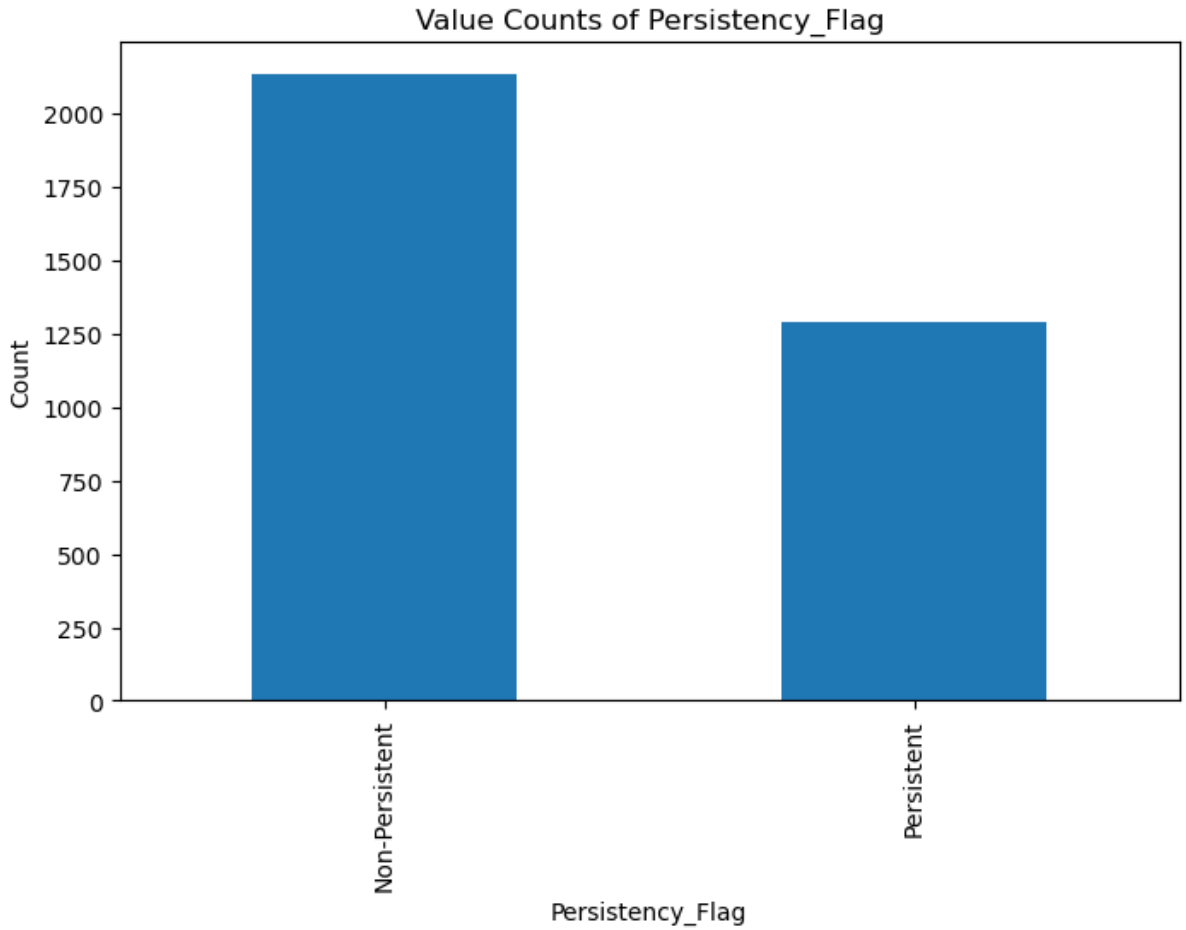
In both groups, the Midwest and South regions have more patients than the other regions.

4) Distribution by Age



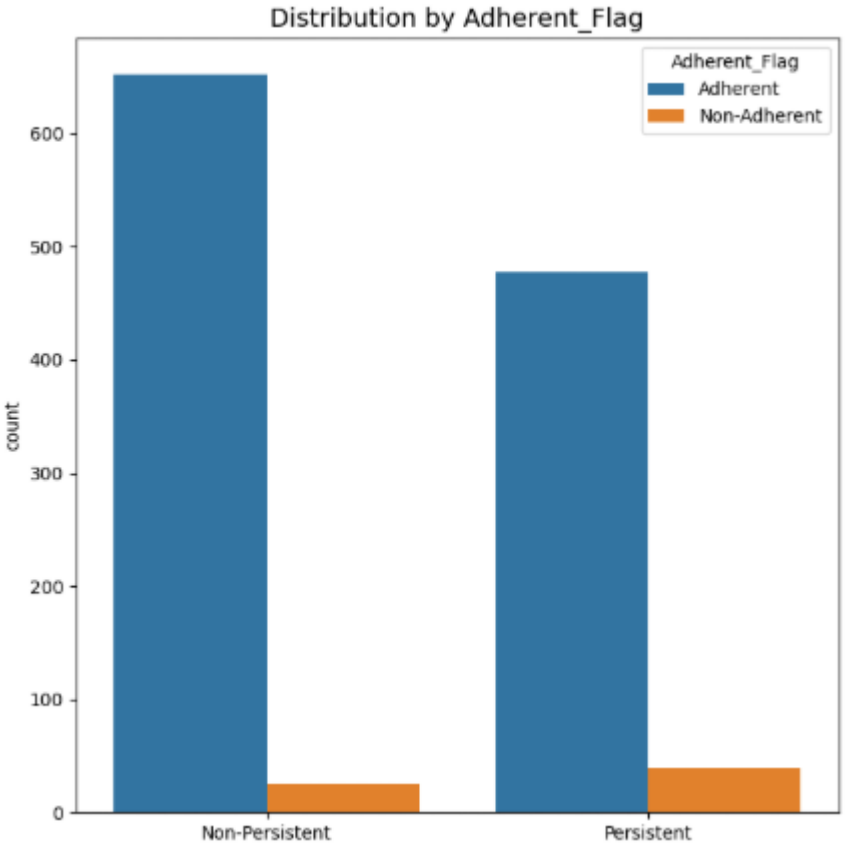
In both groups, 65+ age patients count higher than the others.

5) Non-persistent vs. Persistent



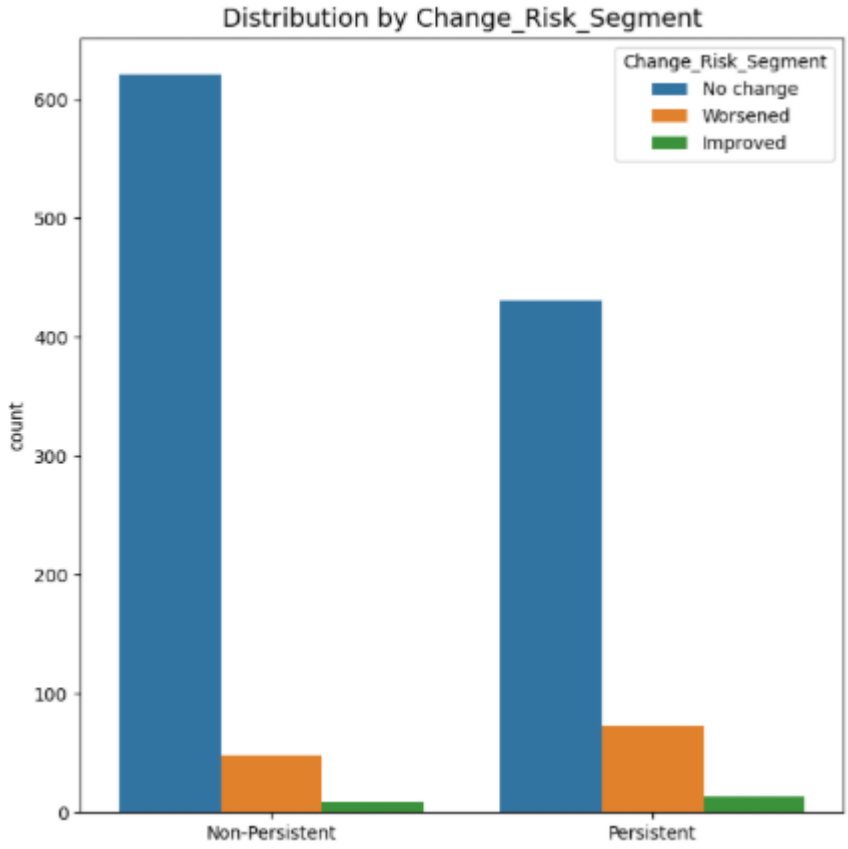
There are more non-persistent patients than the persistent group.

6) Distribution by Adherent_Flag



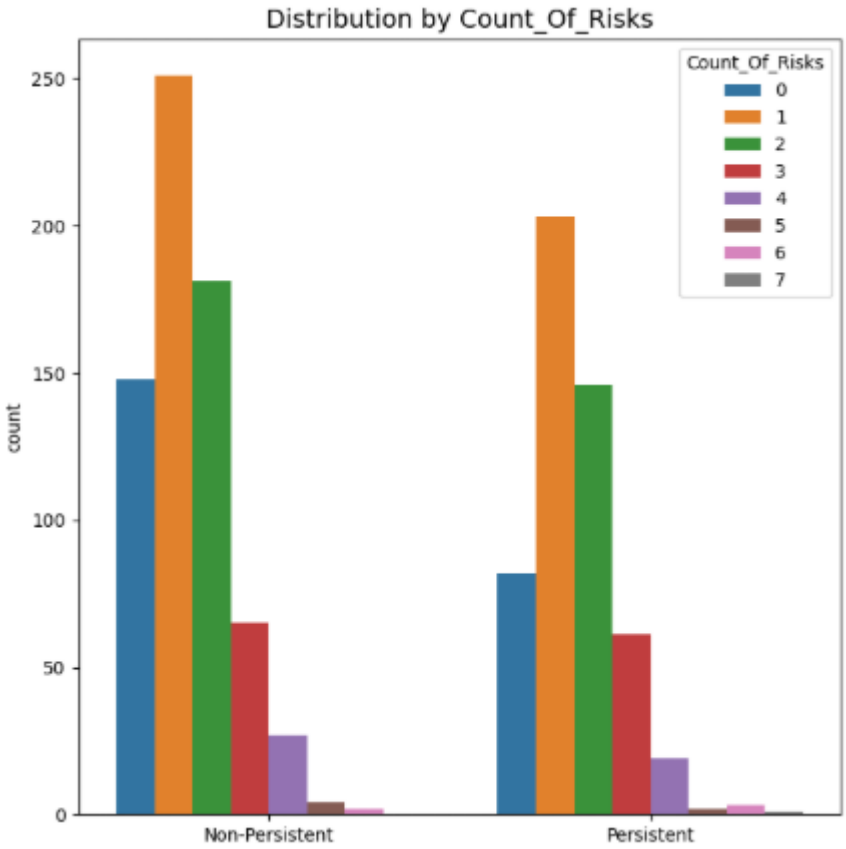
Most of the patients are following their medical advice.

7) Distribution by Change_Risk_Segment



Mostly, there are no risk changes for both groups.

8) Distribution by Count_Of_Risks



Patients with 1 and 2 risks have higher counts than others.

Summary

• Caucasian is the most common race in the study.

• In both groups, females number much more than the number of men.

• In both groups, the Midwest and South regions have more patients than the other regions.

• In both groups, 65+ age patients count higher than the others.

• There are more non-persistent patients than the persistent group.

• Most of the patients are following their medical advice.

• Mostly, there are no risk changes for both groups.

• Patients with 1 and 2 risks have higher counts than others.

Proposed Modeling Technique

Problem Definition

The task involves a supervised binary classification problem where the target variable defines whether a patient is persistent or non-persistent, depending on whether the treatment was followed. The objective of this task is to build a predictive model that can classify patients as either persistent or non-persistent based on demographic, clinical, and behavioral attributes.

Model Selection

This project will implement several machine learning models, such as logistic regression, random forest, and decision trees, to find reliable predictions.

Model Evaluation

Key performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC will be used for model comparisons. The most effective model will be chosen by balancing predictive power and interpretability.