# INTRODUCTION OVERVIEW

- Kaggle’s Space Titanic machine learning competition is quite similar to the well-known Titanic competition. Given a dataset, we are required to predict which passengers were transported or not by an “anomaly” using records recovered from the spaceship’s damaged computer system. The “lore” of the competition is not so important, what you need to know is to develop a machine learning algorithm capable of correctly predicting the outcome of the spaceship’s passengers.

# PROBLEM STATEMENT

While rounding Alpha Centauri en route to its first destination—the torrid 55 Cancri E—the unwary Spaceship Titanic collided with a spacetime anomaly hidden within a dust cloud. Sadly, it met a similar fate as its namesake from 1000 years before. Though the ship stayed intact, almost half of the passengers were transported to an alternate dimension!

To help rescue crews and retrieve the lost passengers, you are challenged to predict which passengers were transported by the anomaly using records recovered from the spaceship’s damaged computer system.

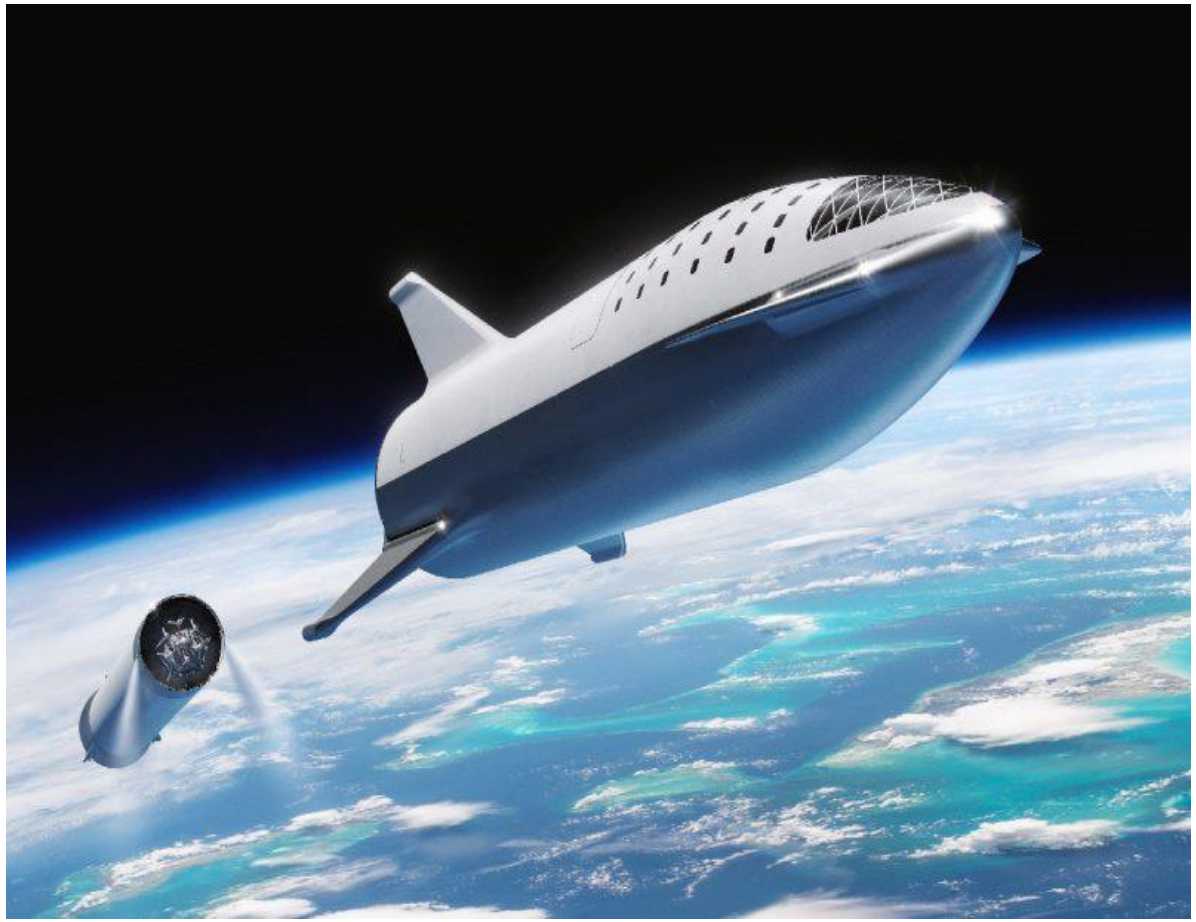
# OBJECTIVES

1. Which HomePlanet had the highest Transportation rate for its passengers.

2. Check whether the passenger being on CryoSleep contributed to the passengers being transported.

3. Examine the trends for the passengers age that were transported to the another dimension.

4. Develop the best machine learning algorith that predicts whether the passengers were transported or not.



# METHODOLOGY

- DATA UNDERSTANDING

Here we get access to all the information about our dataset by examining the statistical analysis as well as missing values and data types

DATA PREPARATION

This involves handling missing values for both numerical and categorical columns.

Handling outliers and duplicate values

- DATA VISUALIZATION

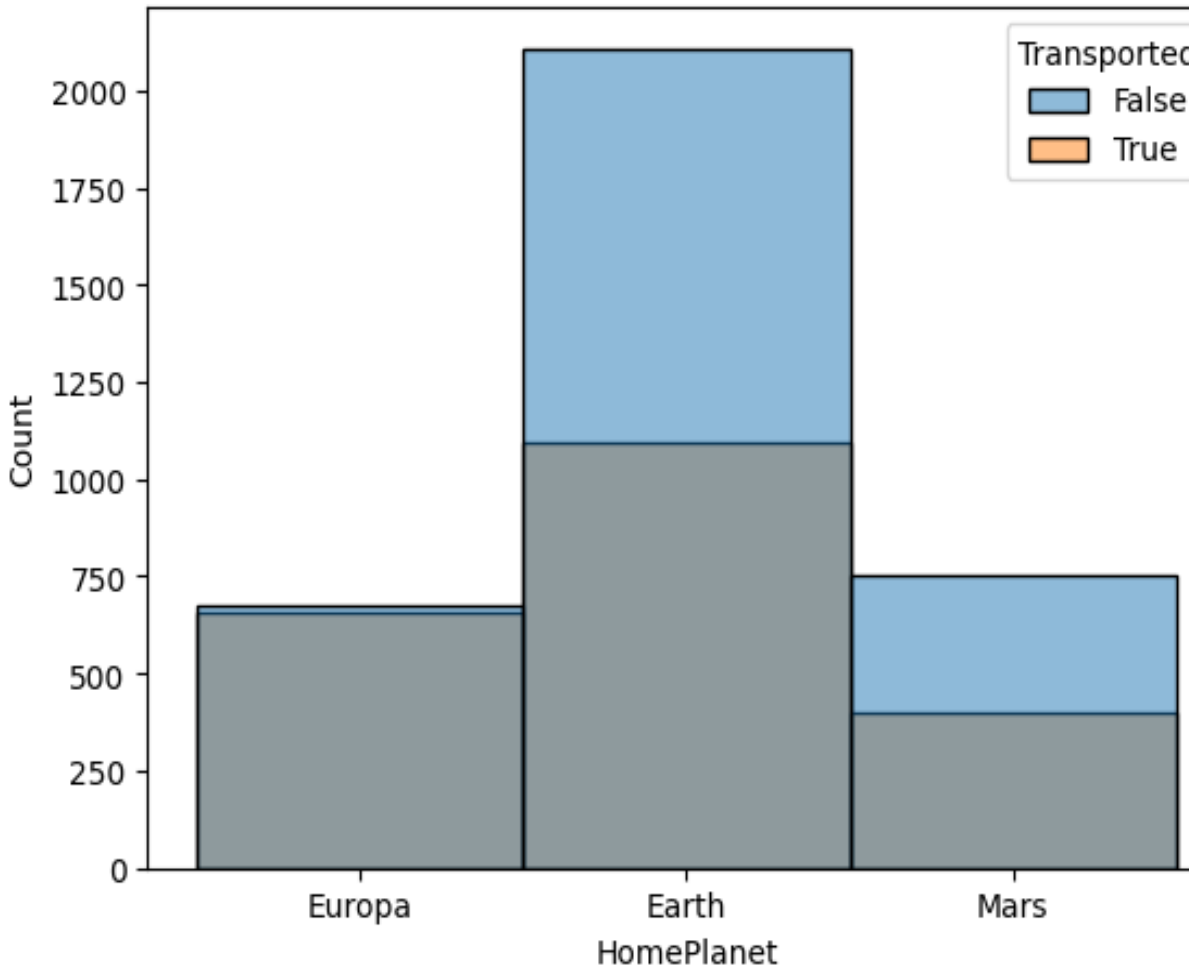
Here we extract insights from the visuals created after univariate and bivariate analysis.

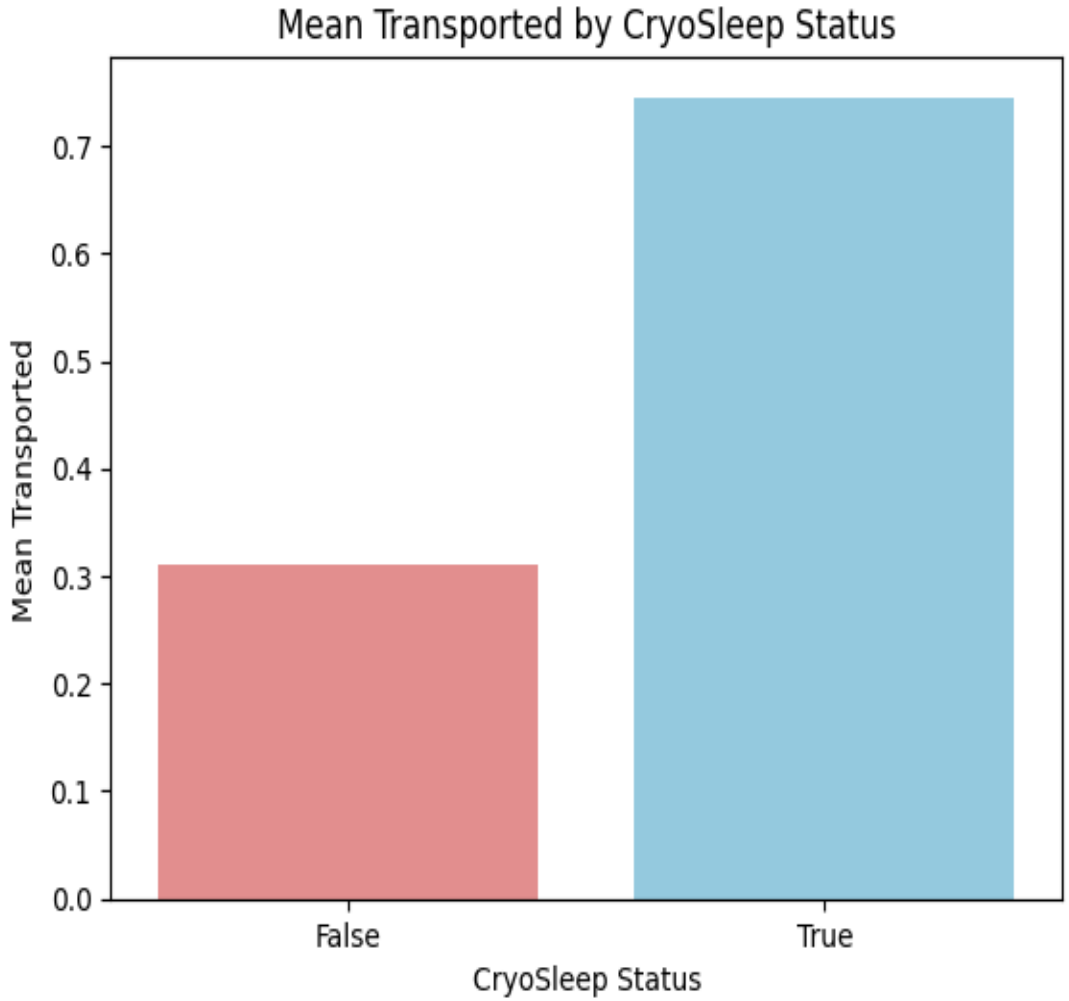
DATA PREPROCESSING

Here select the target and features from our dataset.

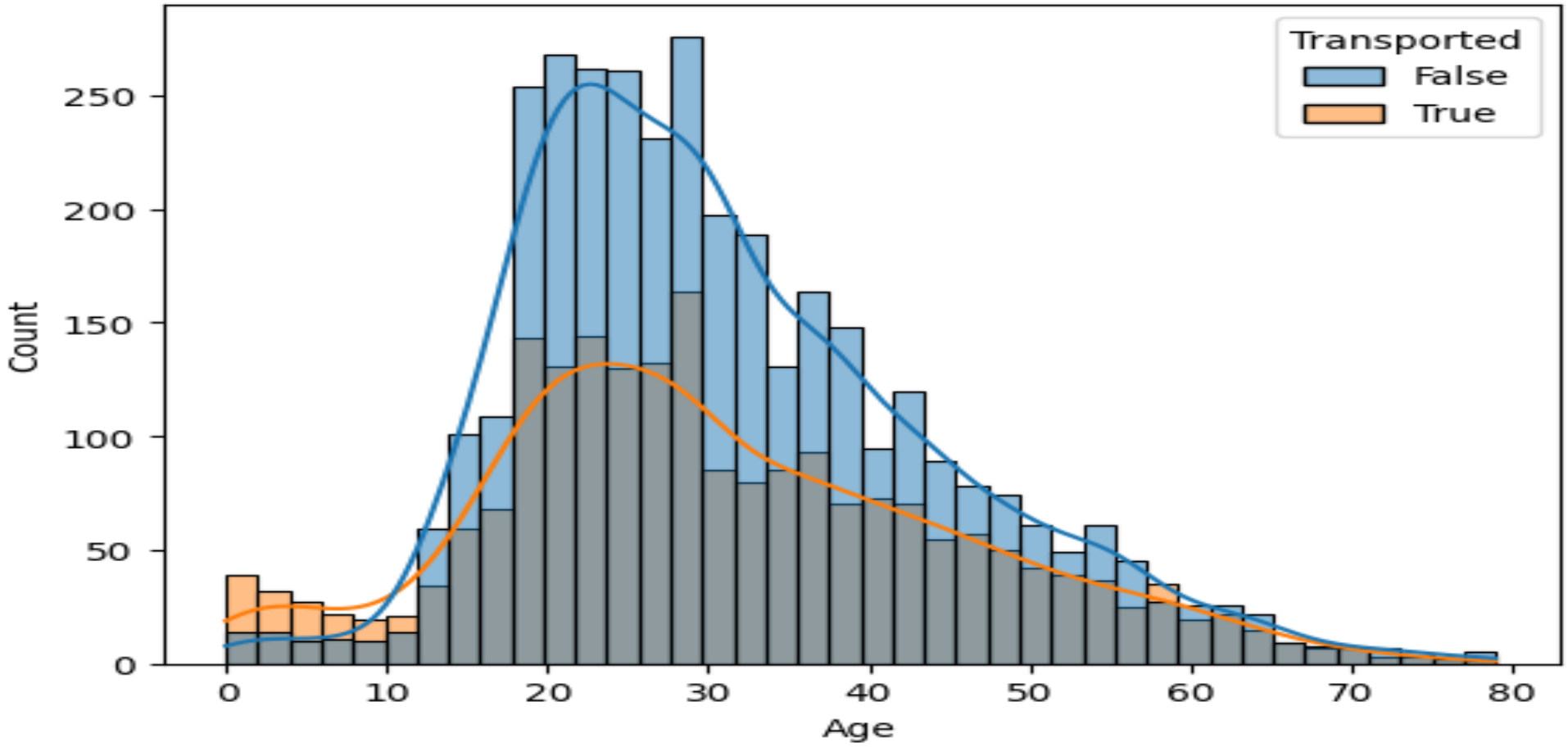
Split the dataset into train and test set.

Perform One Hot Encoding on categorical columns and scaling on the numerical columns.

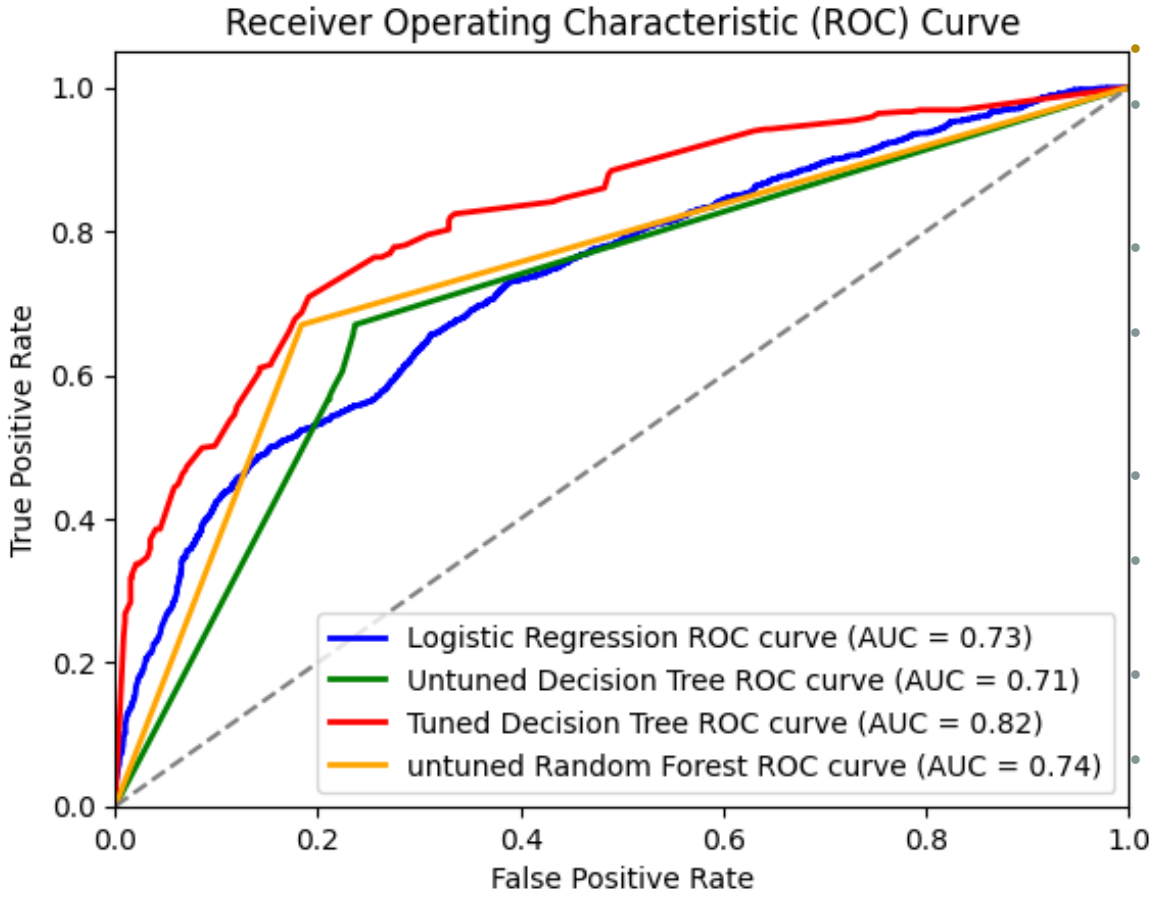

Transported
False
True
2000
1750
1500
1250
1000
750
500
250
0
Count
Europa
Earth
Mars
HomePlanet


Mean Transported by CryoSleep Status
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Mean Transported
False
True
CryoSleep Status

# Visualize the Age of the Passengers and the transportation rate.



# Creating Logistic Regression model, Decision tree and Random Forest




* Tuned Decision Tree (AUC = 0.82):

The tuned decision tree is the best-performing model in terms of distinguishing between the positive and negative classes. It shows significant improvement over the untuned decision tree and other models.

* Untuned Random Forest (AUC = 0.74):

The untuned random forest performs better than the untuned decision tree but slightly better than logistic regression. This shows that random forests, even without tuning, provide decent performance.

* Logistic Regression (AUC = 0.73):

Logistic regression is a relatively simple model, performing slightly worse than the random forest but better than the untuned decision tree.

* Untuned Decision Tree (AUC = 0.71):

The untuned decision tree is the weakest model here, emphasizing the importance of hyperparameter tuning.

# STRATEGIC INSIGHTS

**For Real time Data feed its necessary to** utilize satellites and onboard telemetry to assess the ship's structural location, and environmental conditions.

**AI-Assisted Diagnostics**: Deploy AI systems to analyze damage and predict potential failures.