Cambridge™
INFOTECH
An ISO 9001-2008 Certified Training Institute

INSTITUTE: CAMBRIDGE INFOTECH

NAME: KARTHIK VM

PROJECT REPORT ON: SPACESHIP TITANIC

DOCUMENT TYPE HLD

High-Level Design (HDL)

Spaceship Titanic Analysis Prediction

Table of content

SPACESHIP TITANIC

1. Introduction/Abstract:

Welcome to the year 2912, where your data science skills are needed to solve a cosmic mystery. We've received a transmission from four lightyears away and things aren't looking good.

The Spaceship Titanic was an interstellar passenger liner launched a month ago. With almost 13,000 passengers on board, the vessel set out on its maiden voyage transporting emigrants from our solar system to three newly habitable exoplanets orbiting nearby stars.

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.

Help save them and change history!

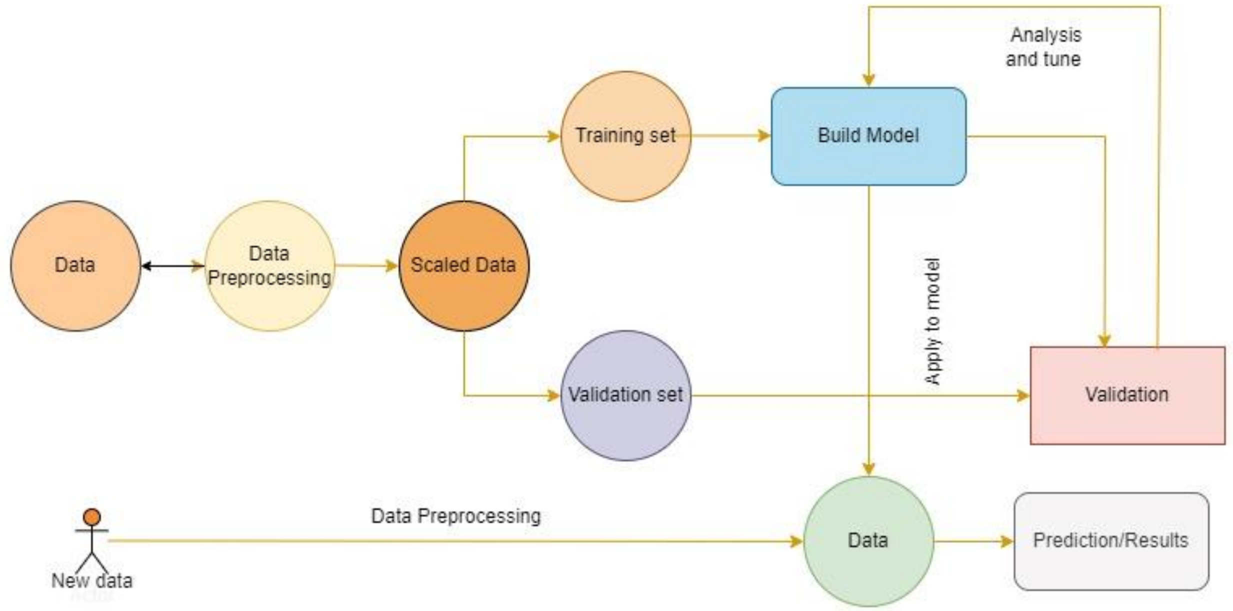
2. Problem Statement (Business Level):

The problem statement of this project is like a spaceship having people from different planets on a voyage but due to some reasons, some people have been transported to another dimension. Our task is to predict who will get transported and who will remain on the spaceship.

3. Problem statement (ML/DL):

1. Objective: To predict the passengers who will get transported and who will remain on the Spaceship Titanic during the voyage using machine learning algorithms
2. Data: The dataset contains information on 8693 passengers, including demographic information PassengerId, Name, Age, Homeplanet, as well as information on the passenger's Cabin, Cryosleep, Destination, VIP, and VIP services such as RoomService, FoodCourt, ShoppingMall, Spa and VRDeck . The dataset also includes a binary variable indicating whether or not the passenger Transported.
3. Evaluation Metric: The model will be evaluated using accuracy as the metric.
4. Constraints: The model should be able to run on a standard laptop with 8GB of RAM and a standard CPU
5. Assumptions: The data is representative of the population of passengers on the Spaceship Titanic and is accurate and complete
6. Expected Results: The model should be able to predict the Transported of passengers on the Spaceship Titanic with an accuracy of at least 75%.

4. Workflow Diagram:



5. Workflow Explanation:

a. Problem definition: Defining the problem of predicting, whether the passengers transported to other dimension or not on the Titanic and identifying the appropriate machine learning techniques to use.

b. Data collection and preprocessing: Collecting and preparing the Spaceship Titanic dataset. This includes tasks such as cleaning, transforming, and normalizing the data.

c. Exploratory data analysis (EDA): Analyzing the data to understand its characteristics and identify patterns and trends. This includes tasks such as visualizing the data, calculating summary statistics and identifying missing values

d. Feature engineering: Creating new features or transforming existing features to improve the performance of the model. This includes tasks such as creating new features based on the existing data, encoding categorical variables, and scaling numerical variables.

e. Model selection and training: Selecting an appropriate machine learning model and training it on the data. This includes tasks such as selecting the appropriate algorithm, tuning the model's parameters, and evaluating the model's performance.

f. Model evaluation: Evaluating the performance of the model using appropriate evaluation metrics such as accuracy, precision, recall and f1-score.

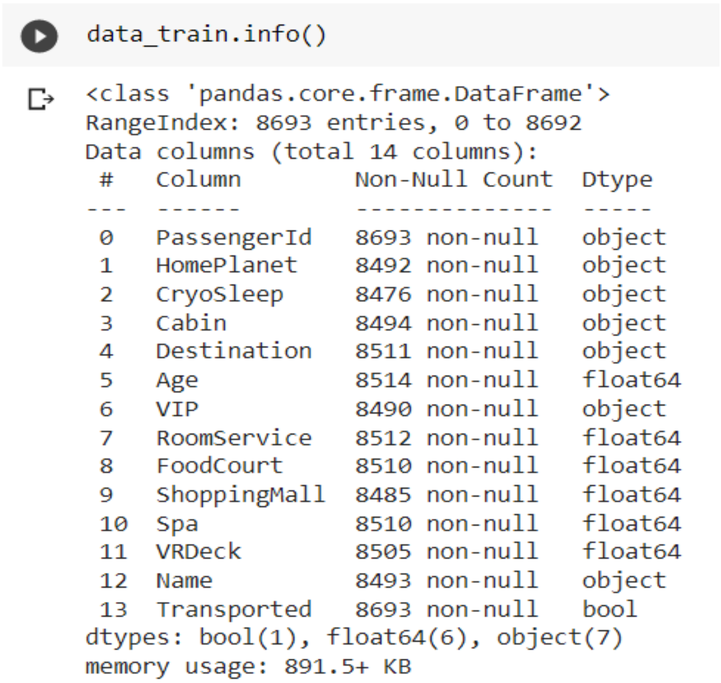
g. Model Tuning: Try to reduce the overfitting or underfitting on the select model and get the best estimator and best parameters. And Create a new model with the best estimator and predict

h. Final Prediction on the Test (Unseen data): After identifying the best hyperparameter-tuned model, now predict the results from the finalized model

6. DATASET DESCRIPTION:

Features:

- PassengerId - A unique Id for each passenger. Each Id takes the form gggg_pp where gggg indicates a group the passenger is traveling with and pp is their number within the group. People in a group are often family members, but not always.
- HomePlanet - The planet the passenger departed from, typically their planet of permanent residence.
- CryoSleep - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
- Cabin - The cabin number where the passenger is staying. Takes the form deck/num/side, where side can be either P for Port or S for Starboard.
- Destination - The planet the passenger will be debarking to.
- Age - The age of the passenger.
- VIP - Whether the passenger has paid for special VIP service during the voyage.
- RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Amount the passenger has billed at each of the Spaceship Titanic's many luxury amenities.
- Name - The first and last names of the passenger.
- Transported - Whether the passenger was transported to another dimension. This is the target, the column you are trying to predict.



```
data_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8693 entries, 0 to 8692
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   PassengerId   8693 non-null   object
 1   HomePlanet    8492 non-null   object
 2   CryoSleep     8476 non-null   object
 3   Cabin         8494 non-null   object
 4   Destination   8511 non-null   object
 5   Age           8514 non-null   float64
 6   VIP           8490 non-null   object
 7   RoomService   8512 non-null   float64
 8   FoodCourt     8510 non-null   float64
 9   ShoppingMall  8485 non-null   float64
 10  Spa           8510 non-null   float64
 11  VRDeck        8505 non-null   float64
 12  Name          8493 non-null   object
 13  Transported   8693 non-null   bool
dtypes: bool(1), float64(6), object(7)
memory usage: 891.5+ KB
```

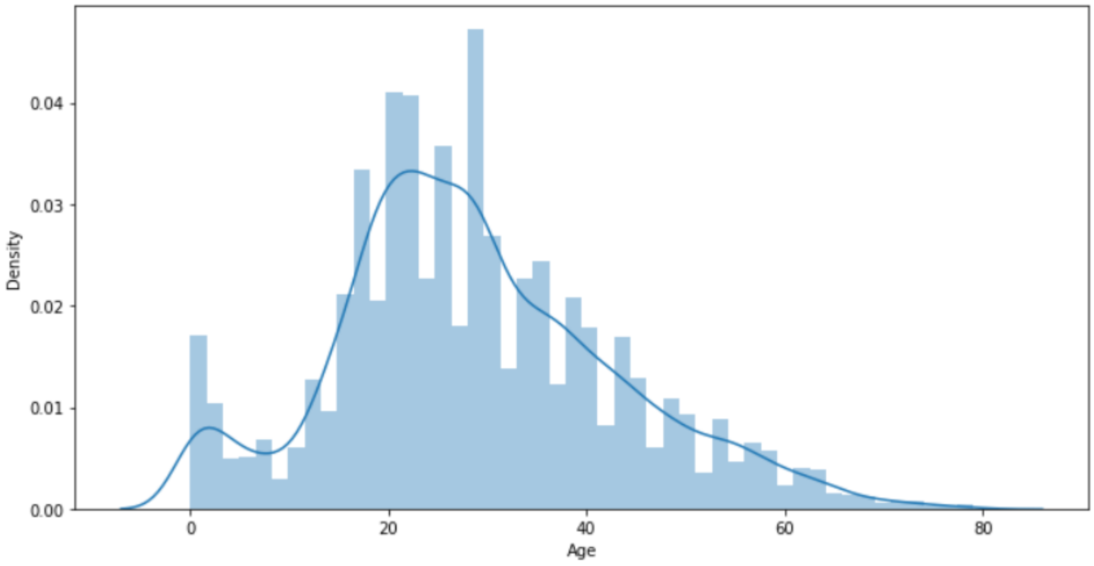
In this we get the total number of columns, name of the columns and the Dtypes I,e. Data type, so we have total 14 columns which has 13 independent variable and 1 is dependent variable and also we got to know that we have 1 bool, 6 floating datatype and 7 object type.

7. EDA RESULTS:

Univariate Analysis:

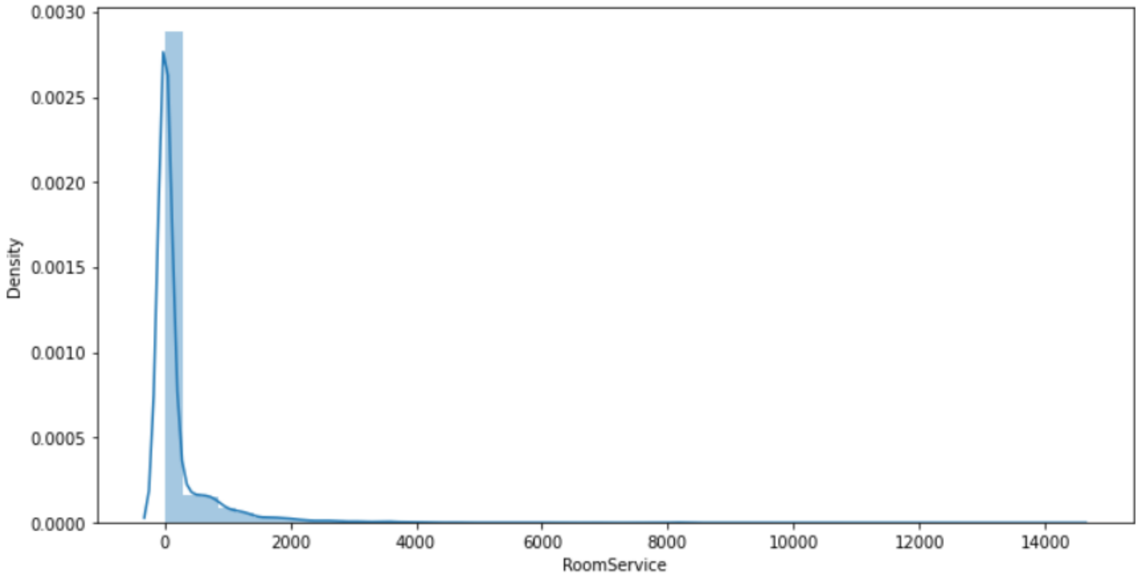
Let's is the numerical distribution in the dataset.

a) Age distribution:



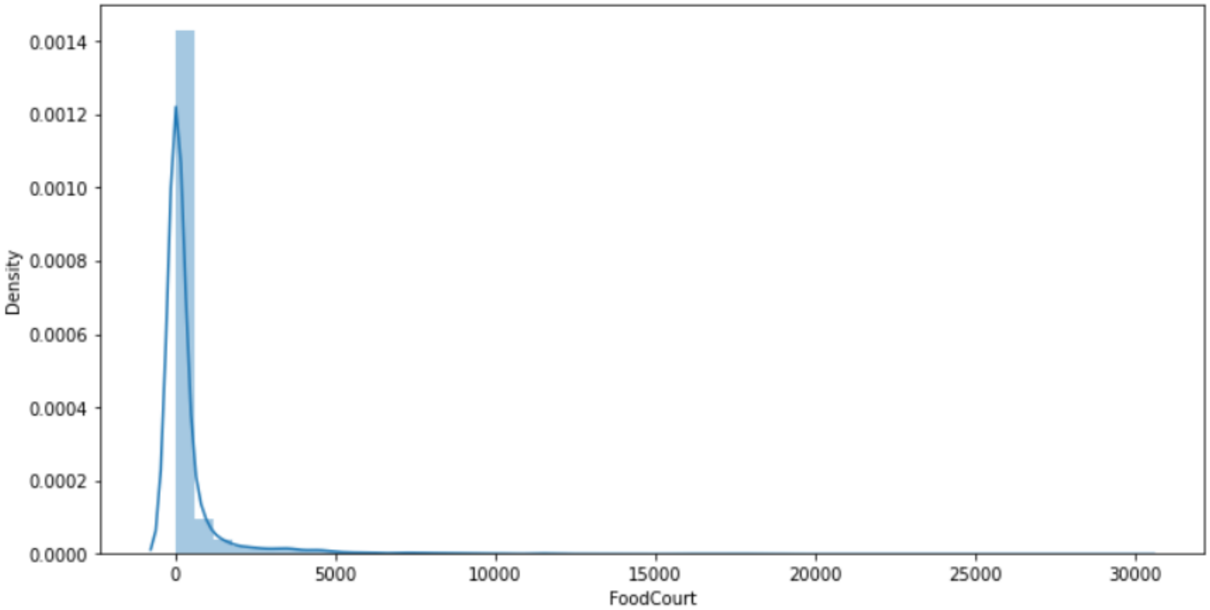
It is similar to normal distribution or gaussian distribution

b) Room Service:



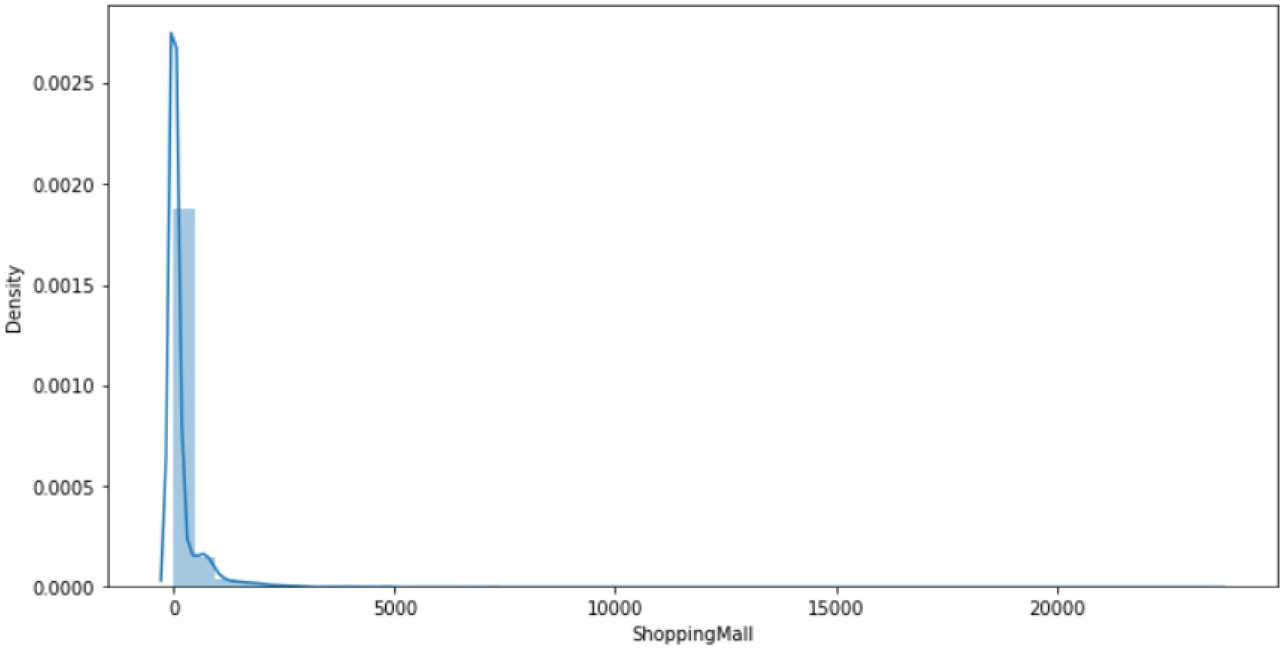
It is highly right Skewed data where we need to make it has normal distribution in the future steps

c) Food Court



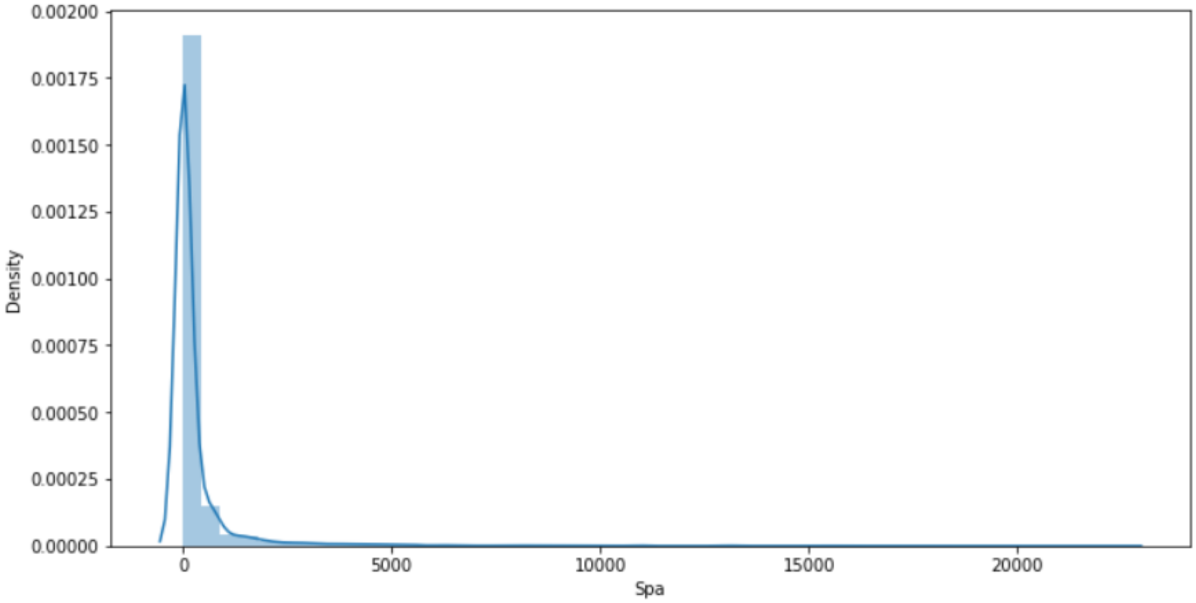
It is highly right Skewed data where we need to make it has normal distribution in the future steps

d) Shopping Mall



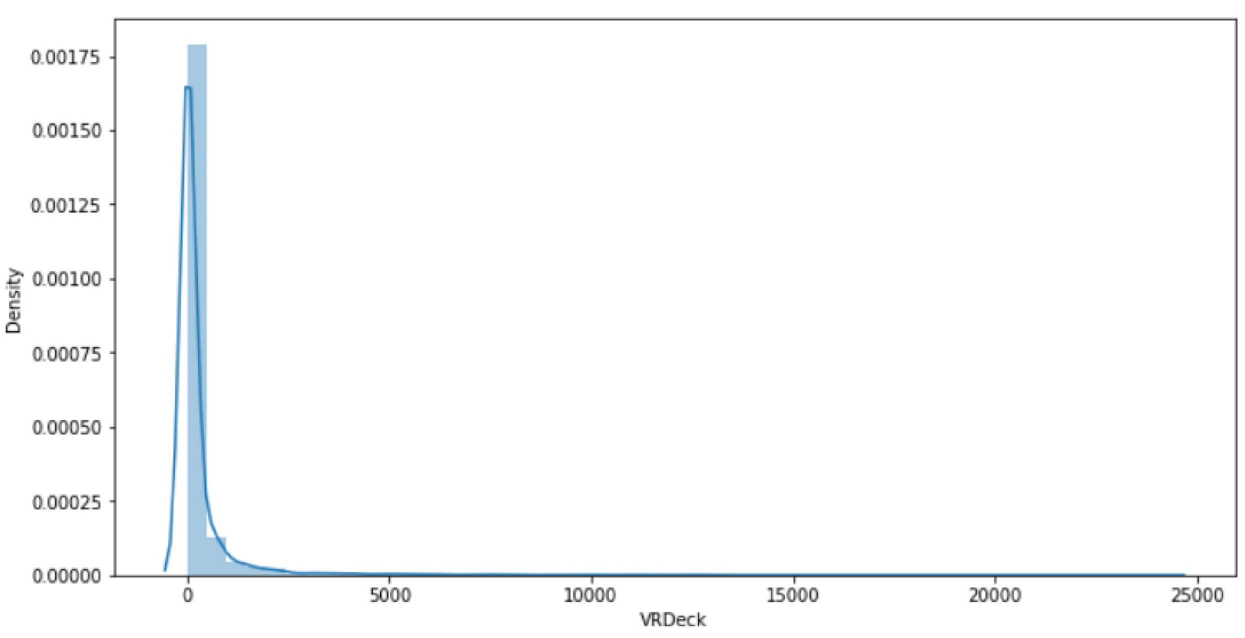
It is highly right Skewed data where we need to make it has normal distribution in the future steps

e) Spa



It is highly right Skewed data where we need to make it has normal distribution in the future steps

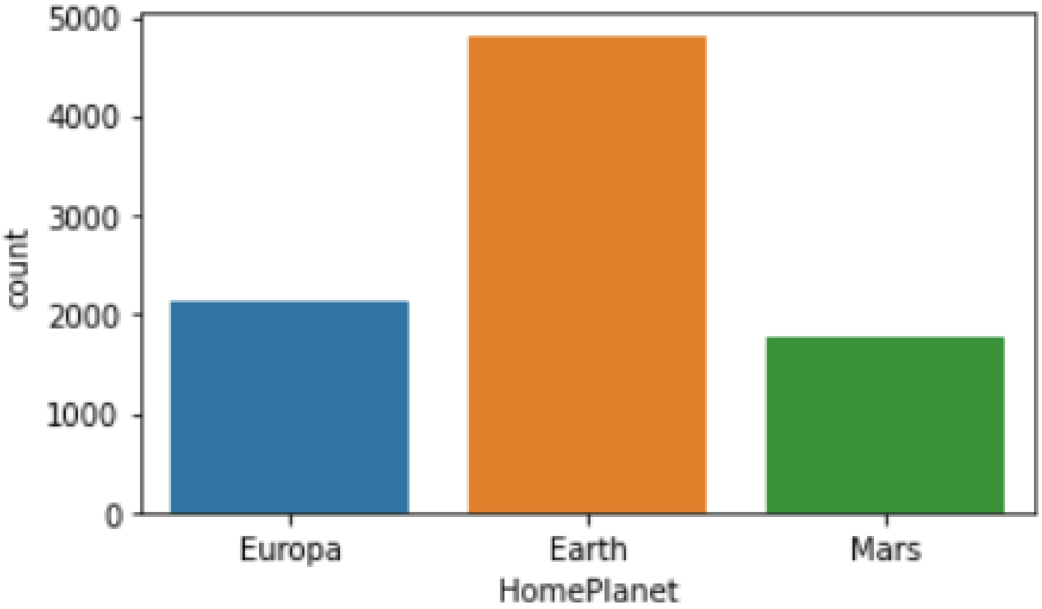
f) VRDeck



It is highly right Skewed data where we need to make it has normal distribution in the future steps

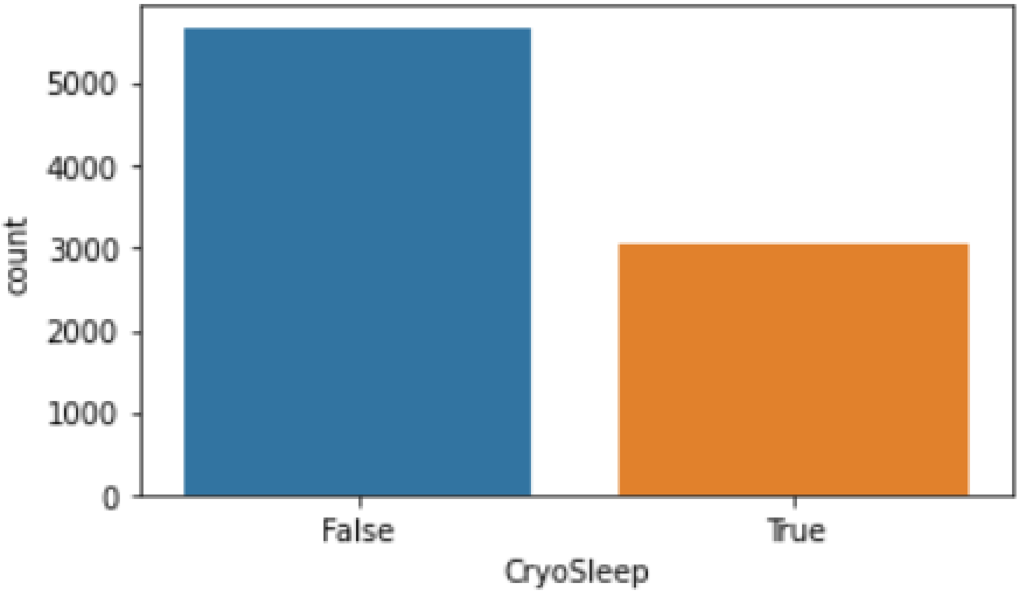
Categorical Variables distributions:

a) Home Planet



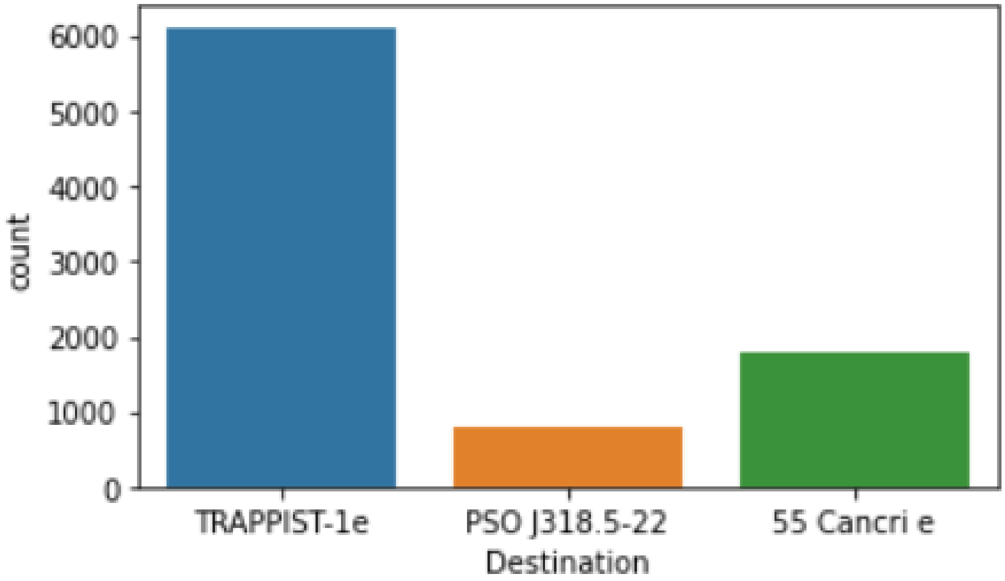
From the above graph, we can say that the majority of the people transported are from the Planet Earth when compared to Europa and Mars

b) Cryo Sleep



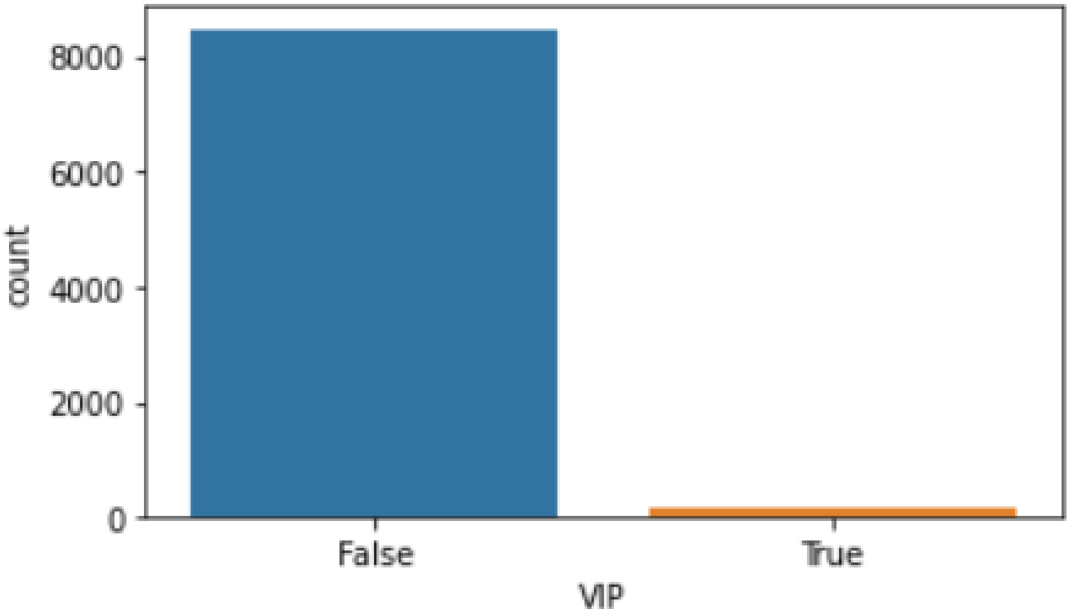
From the above graph, we can say that Here we can observe that those who are in CryoSLeep have higher chances of getting transported

c) Destination



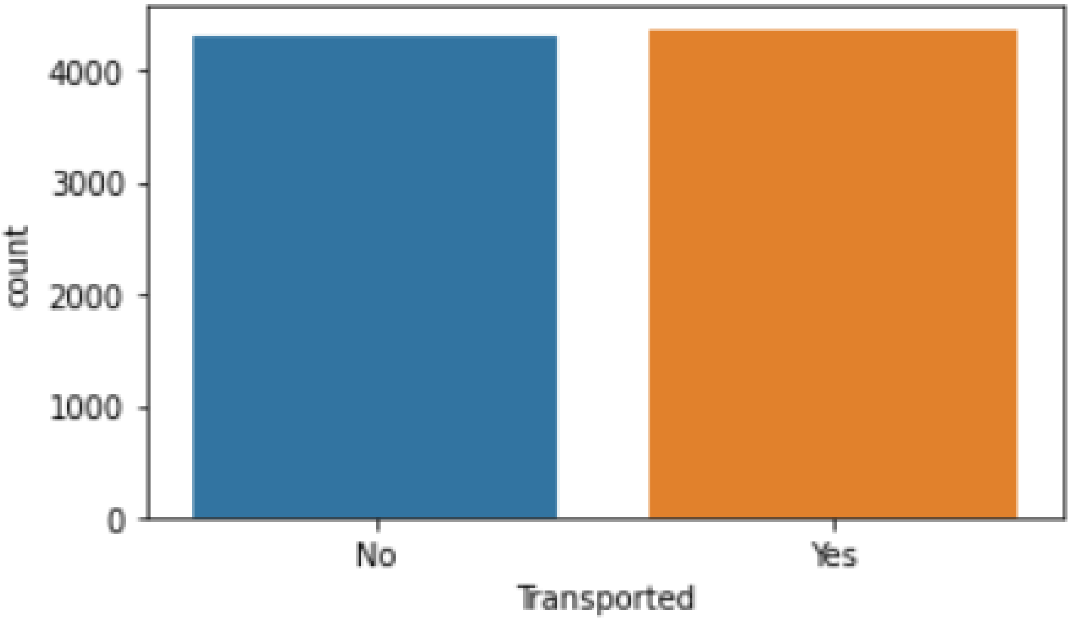
Here we can observe that the majority of the transported to the destination TRAPPIST-1e.

d) VIP



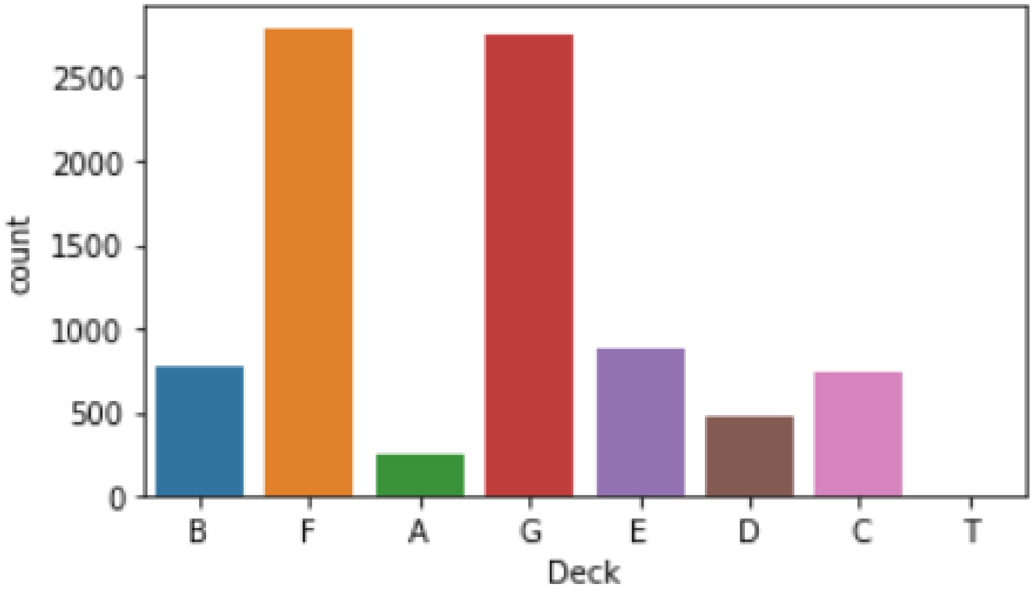
Here the majority of the people during the voyage not opt for the VIP service.

e) Transported



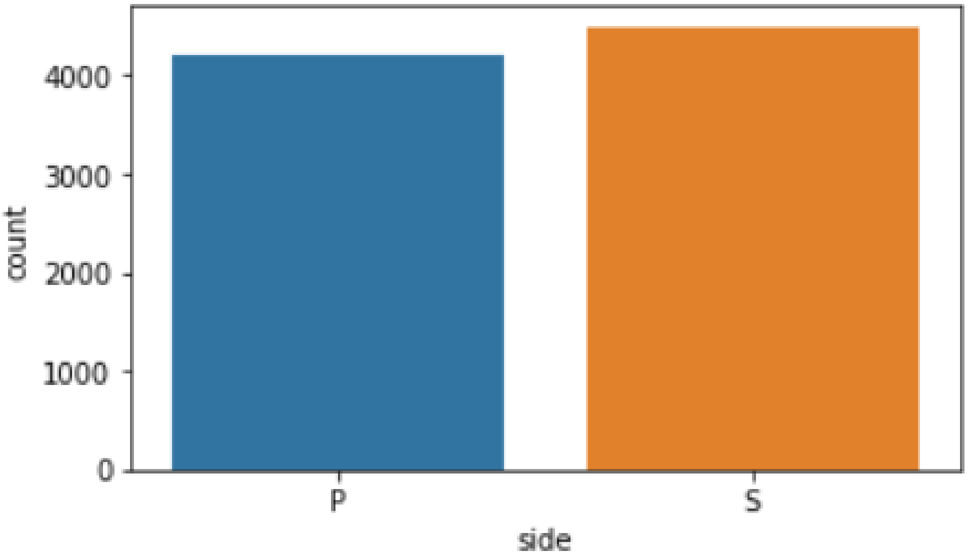
Here we can observe that the people transported during the voyage is almost equal

f) Deck



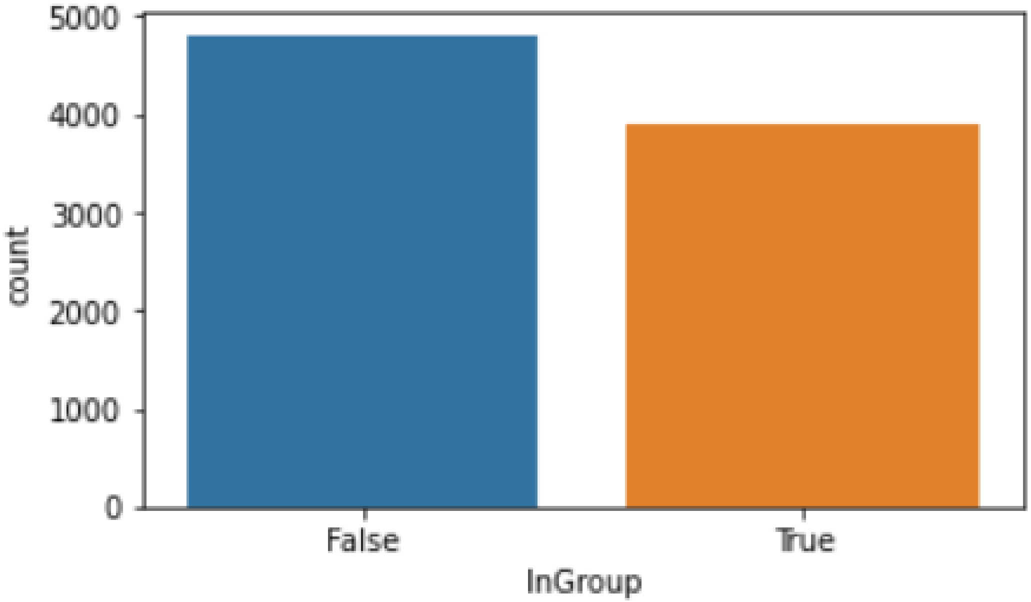
The decks a variable which is separated as a new column from the VRDeck, here we can observe that from deck F and G we have more number of people.

g) Side



Side also the new variable which is separated from the VRDeck, so here we can say that the people equal opt for the P for Port or S for Starboard.

h) InGroup

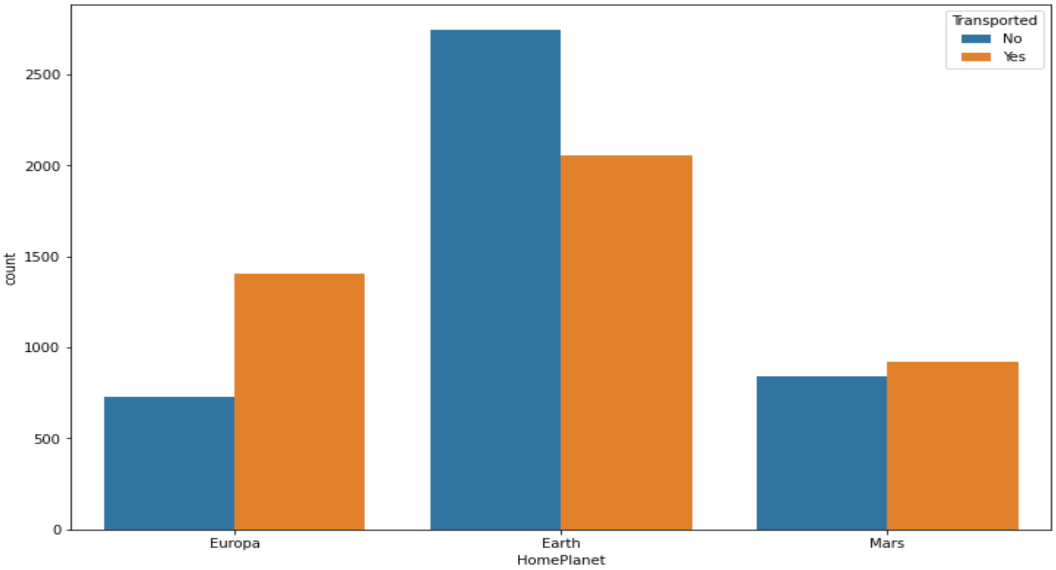


Ingroup is a variable which is created from the GroupId, the majority of the people are not from the same group.

Bivariate Analysis:

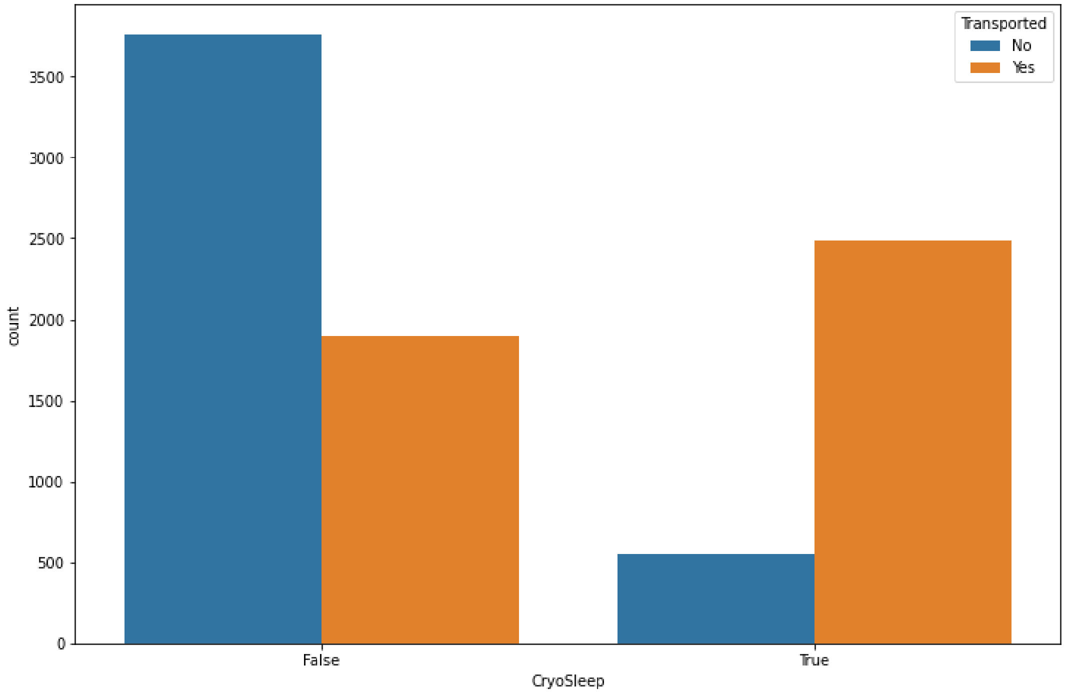
Categorical variable vs Traget(Transported)

a) Home planet vs Transported



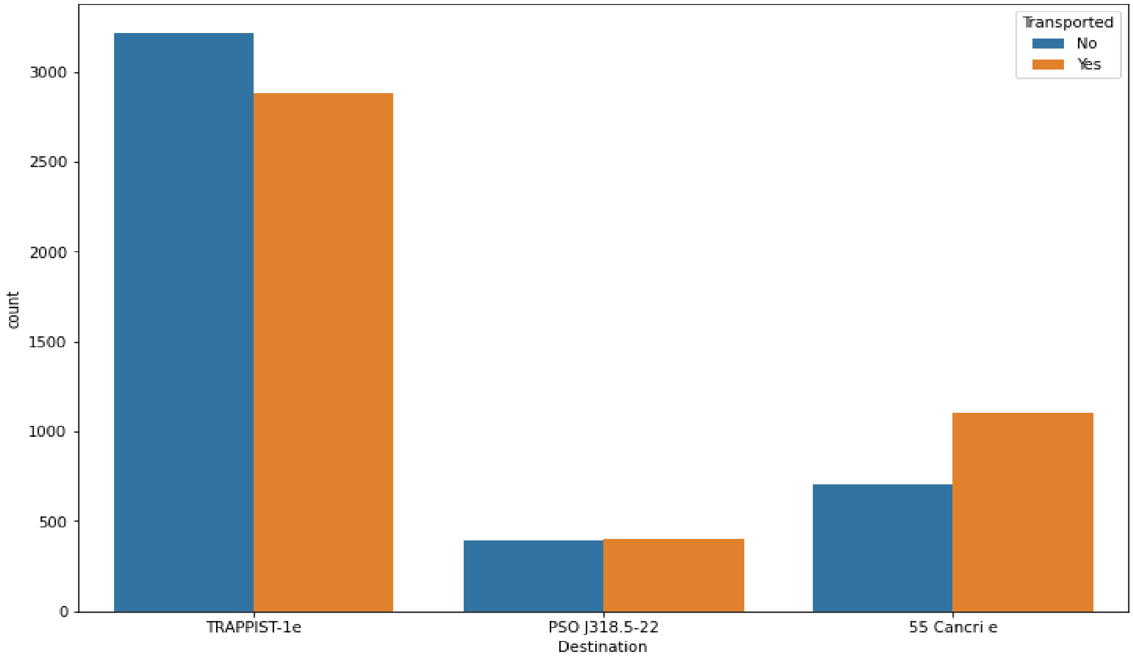
Here we can see the individual distribution of Home Planet according to the target column that is Transported or Not.

b) CryoSleep vs Transported



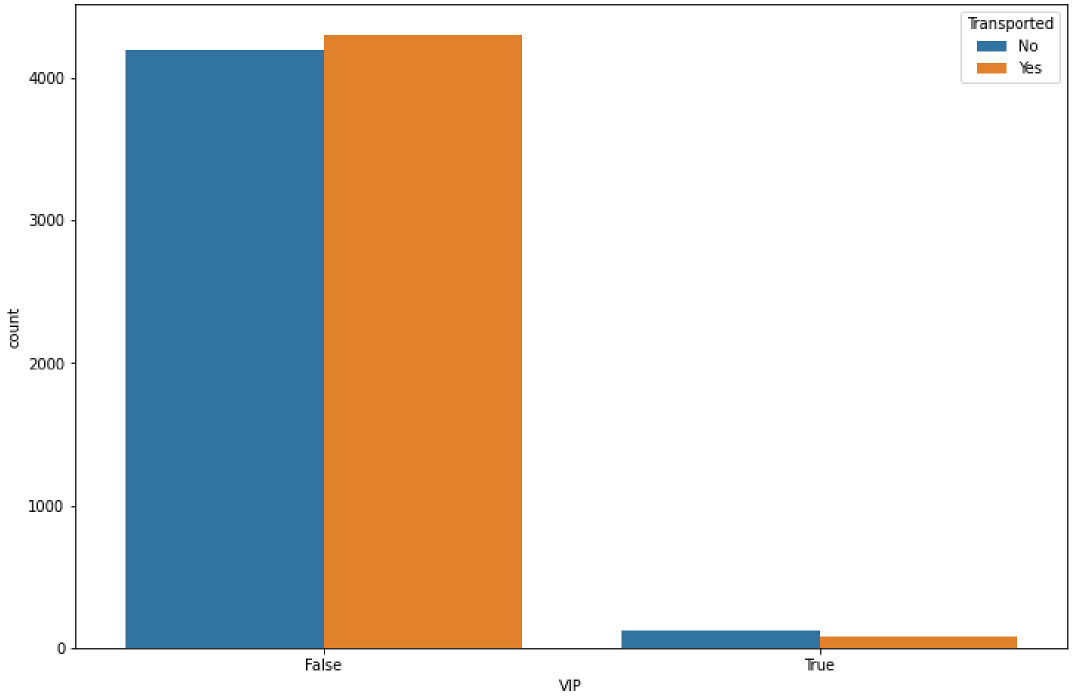
Here we can see the individual distribution of Cryosleep according to the target column that is Transported or Not.

c) Destination Vs Transported



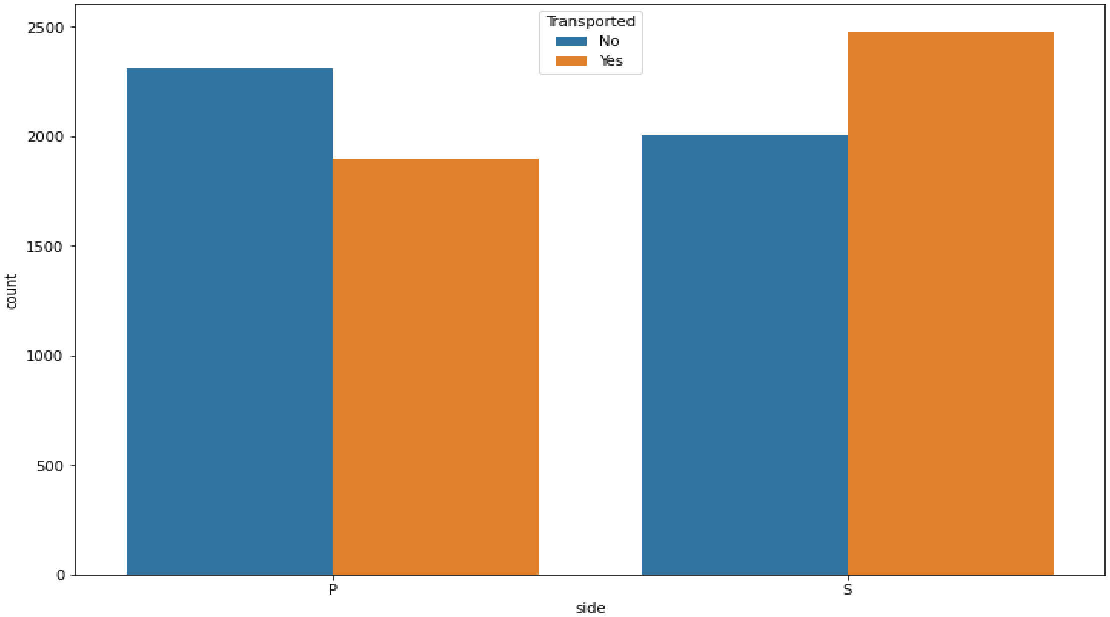
Here we can see the individual distribution of Destination according to the target column that is Transported or Not.

d) VIP vs Transported



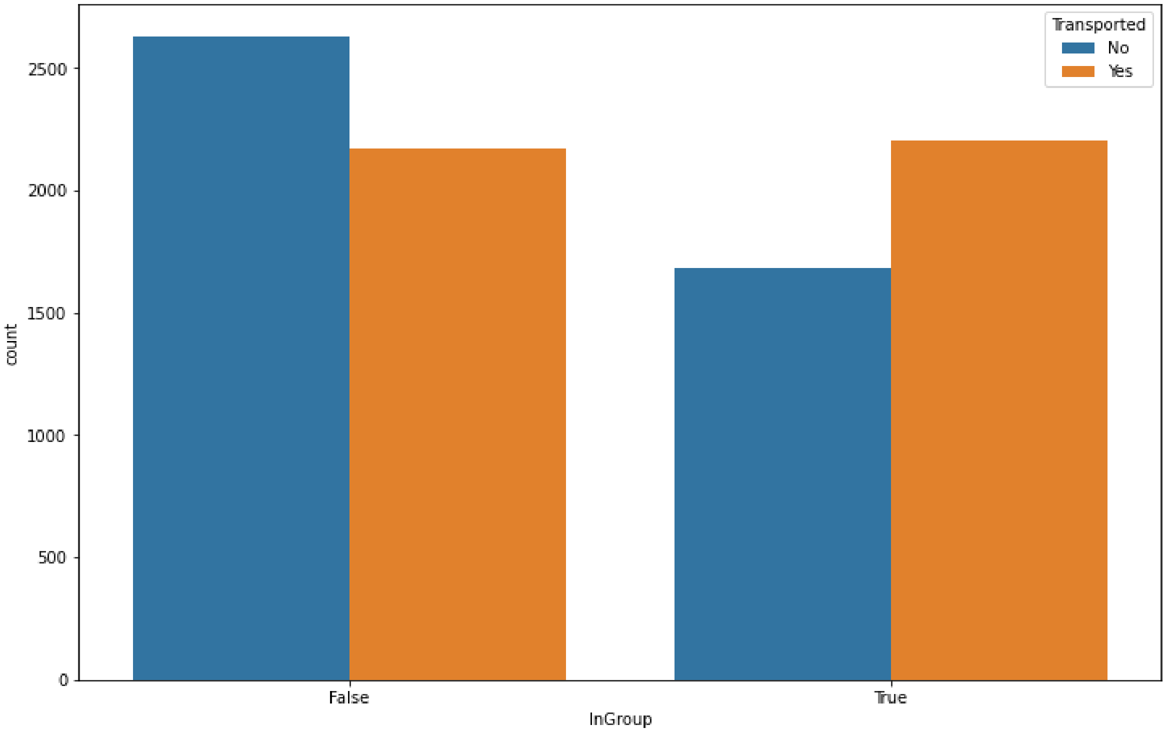
Here we can see the individual distribution of VIP according to the target column that is Transported or Not.

e) Side vs Transported



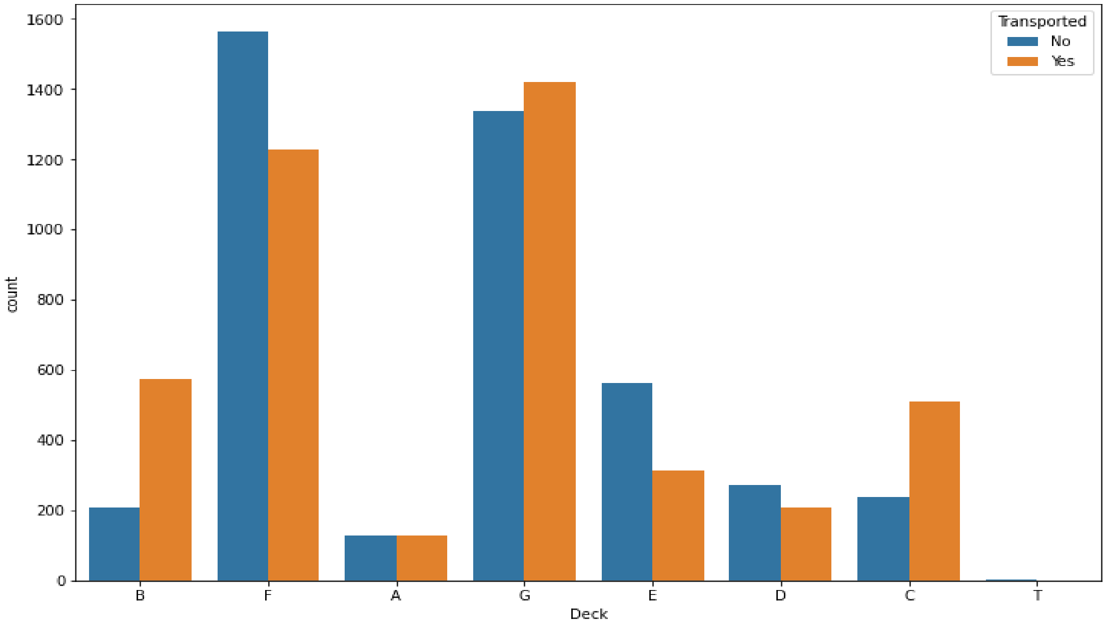
Here we can see the individual distribution of Home Planet according to the target column that is Transported or Not.

f) InGroup Vs Transported



Here we can see the individual distribution of InGroup according to the target column that is Transported or Not.

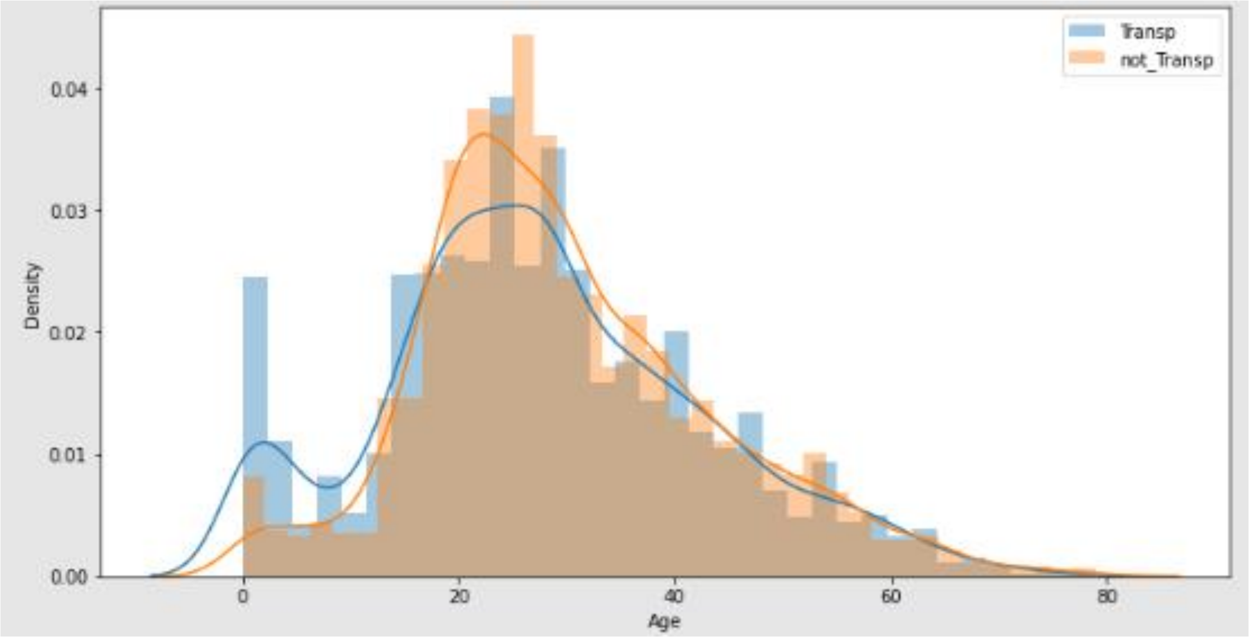
g) Deck Vs Transported



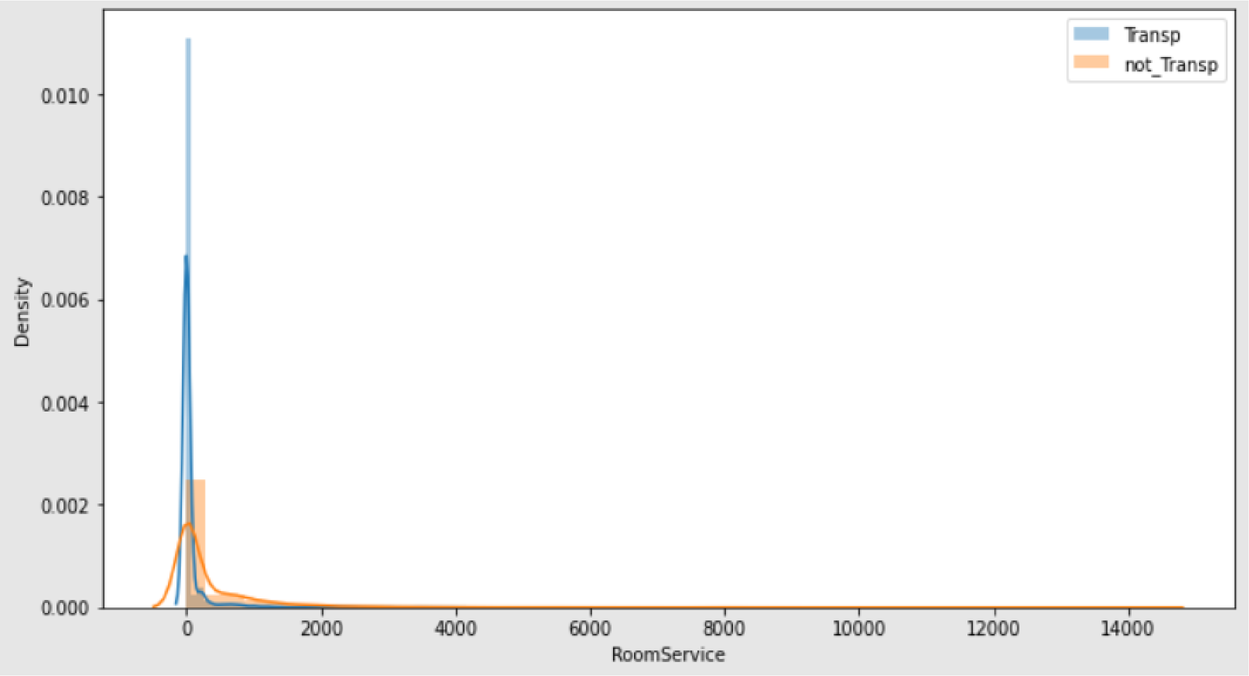
Here we can see the individual distribution of Deck according to the target column that is Transported or Not.

Numerical Variable with Target (Transported)

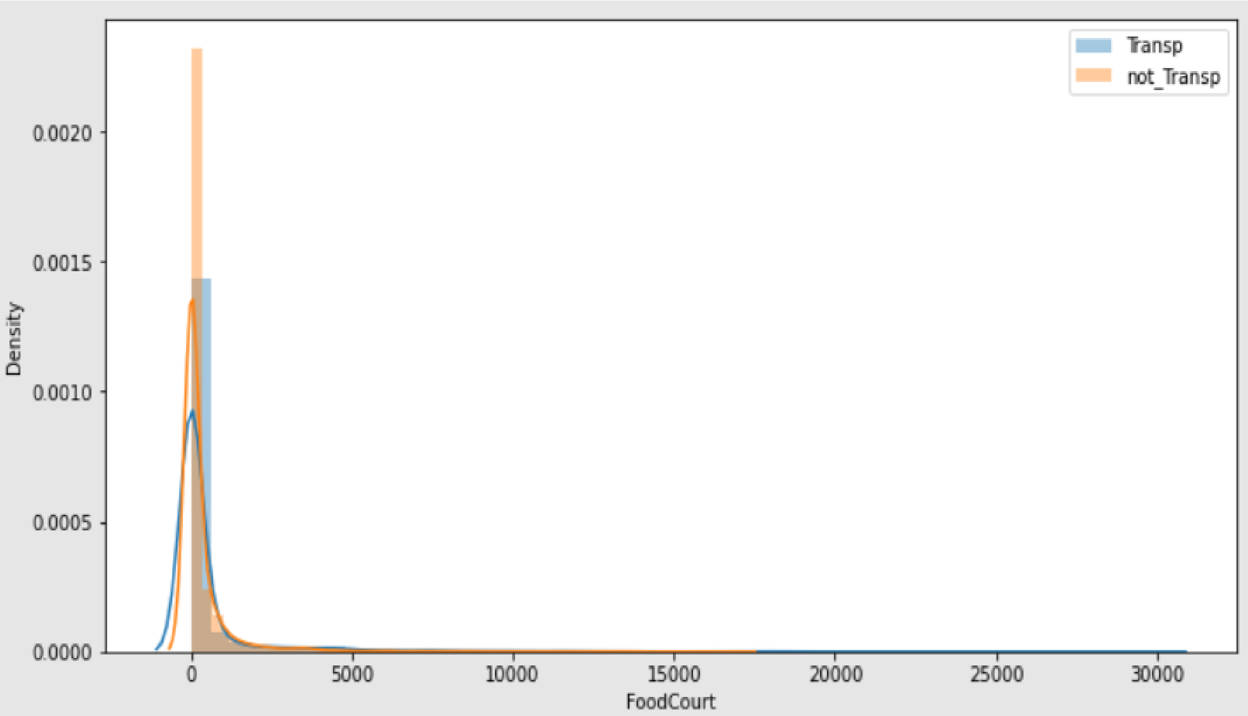
a) Age Vs Transported



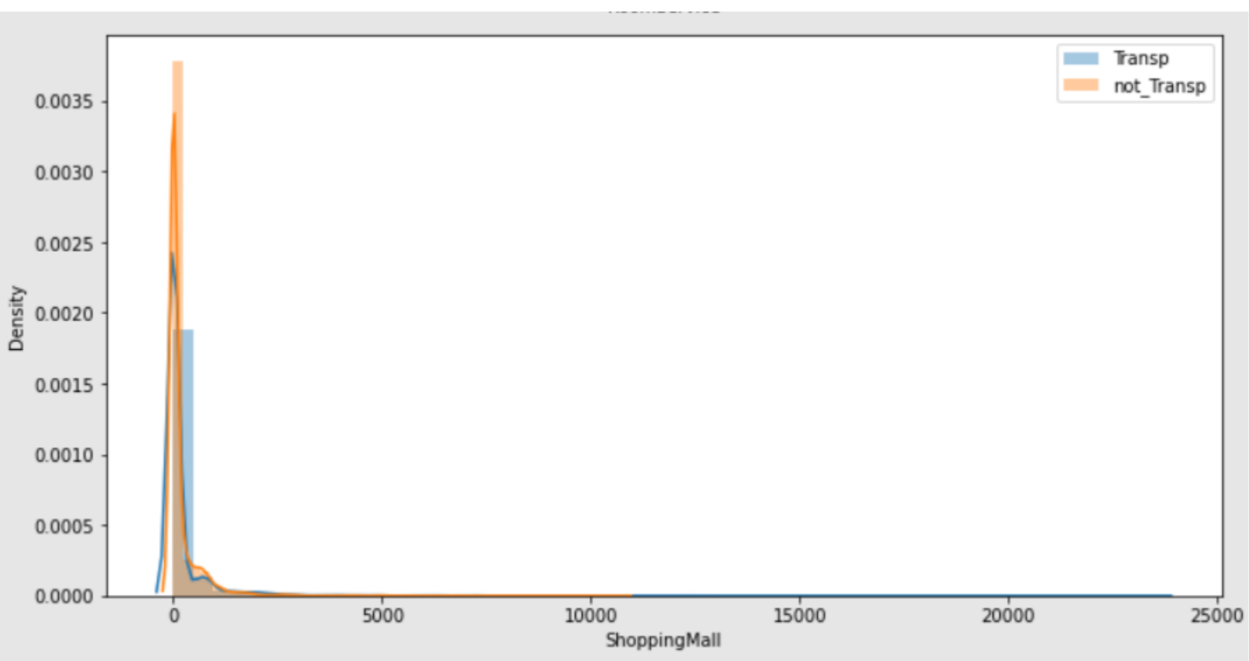
b) Room Service vs Transported



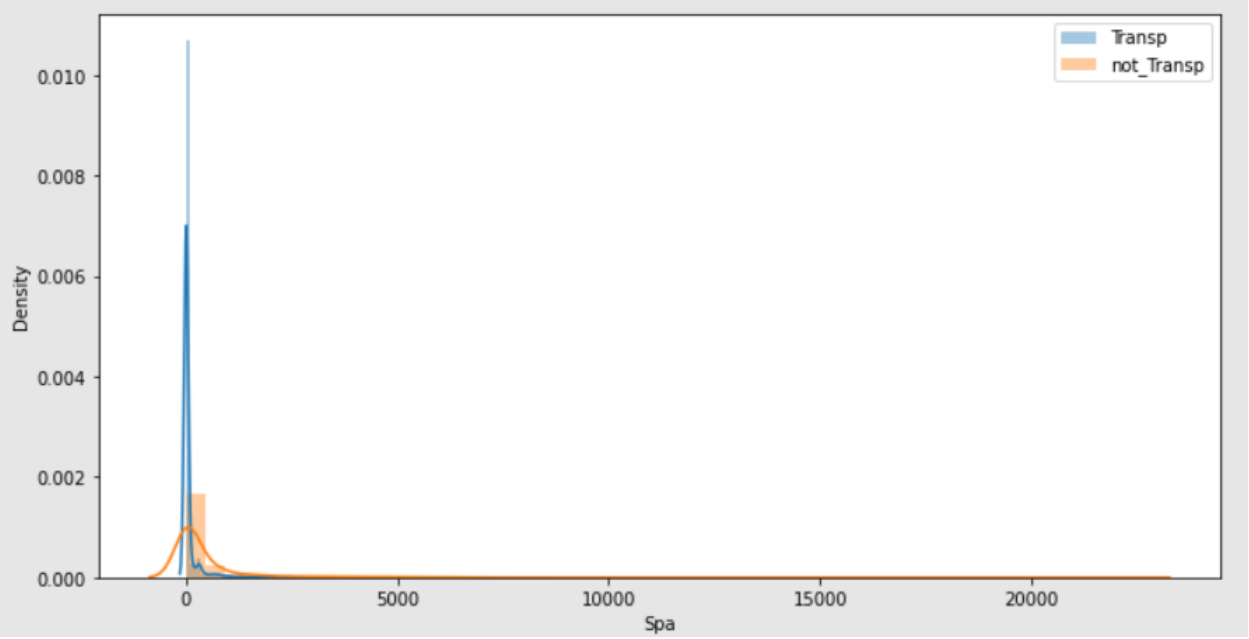
c) Food Court Vs Transported



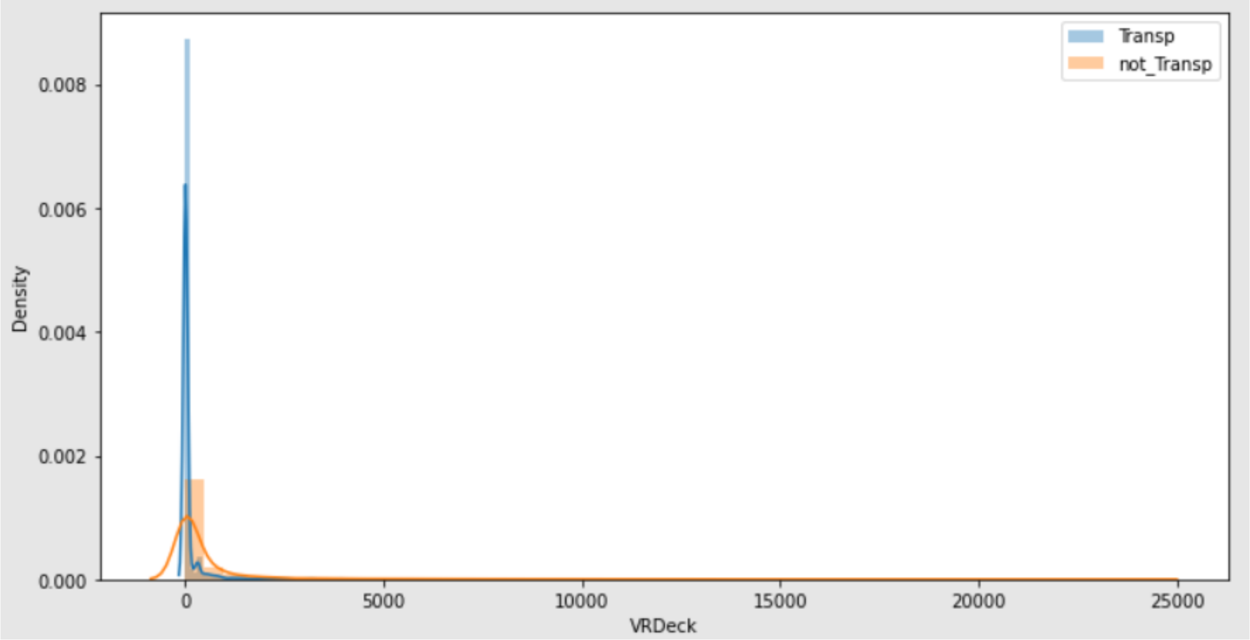
d) Shopping Mall vs Transported



e) Spa vs Transported



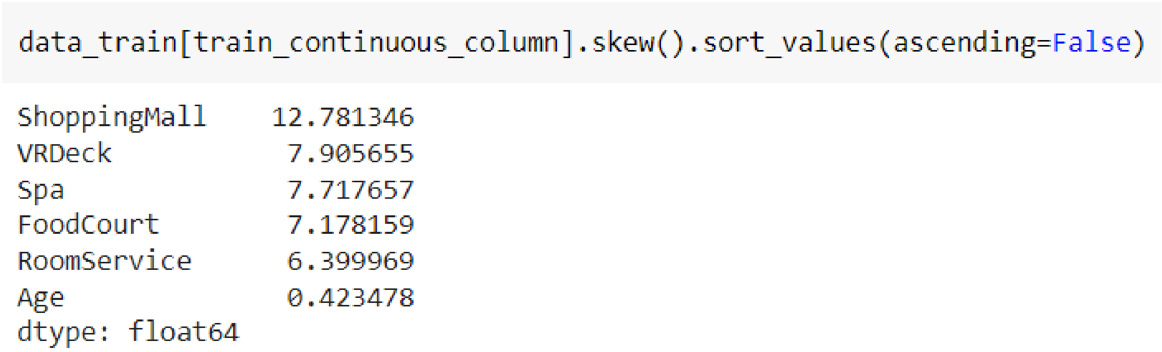
f) VRDeck Vs Transported



8. Outliers:

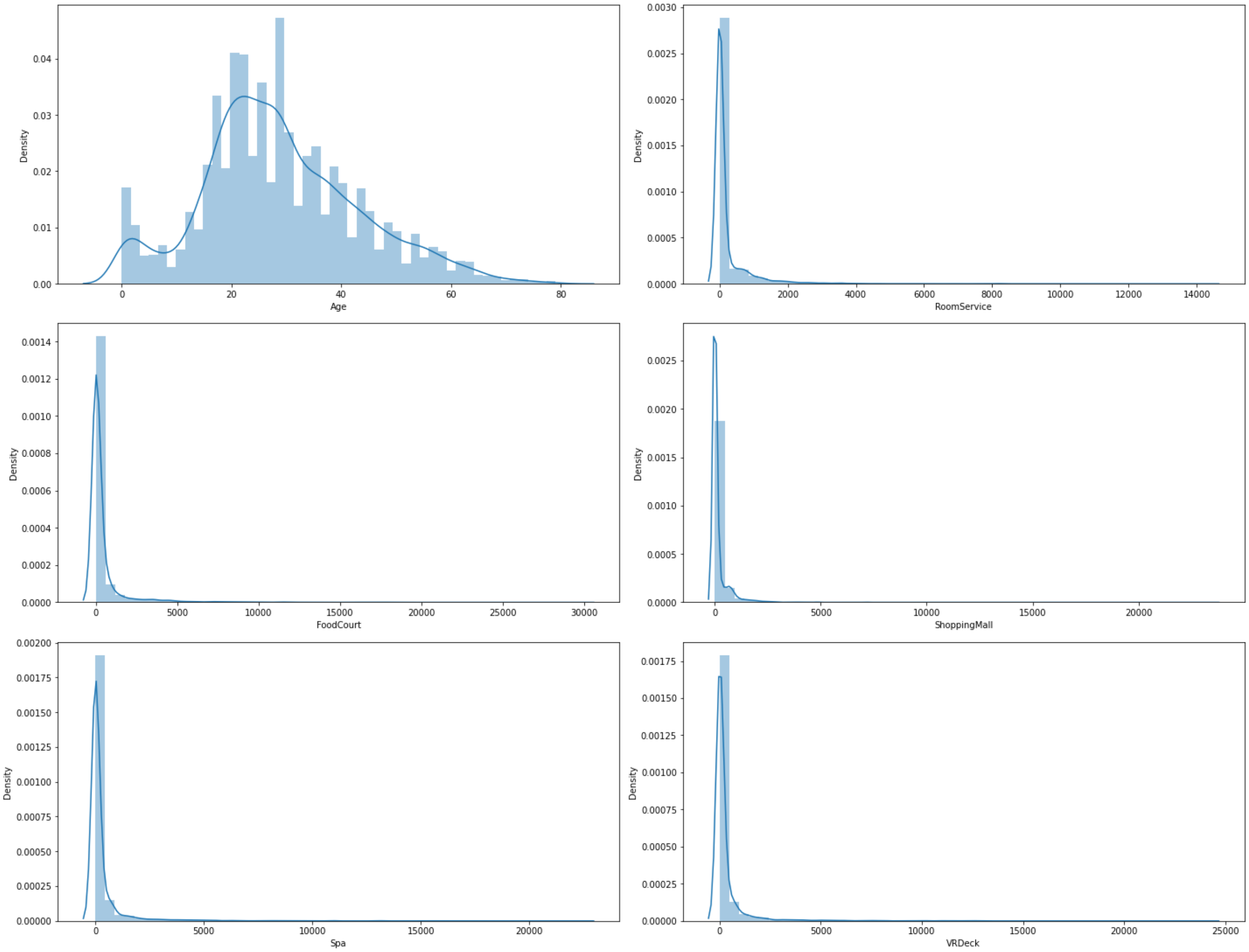
Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

In this dataset, we have right skewed data



```
data_train[train_continuous_column].skew().sort_values(ascending=False)
```

```
ShoppingMall    12.781346
VRDeck           7.905655
Spa              7.717657
FoodCourt        7.178159
RoomService      6.399969
Age              0.423478
dtype: float64
```

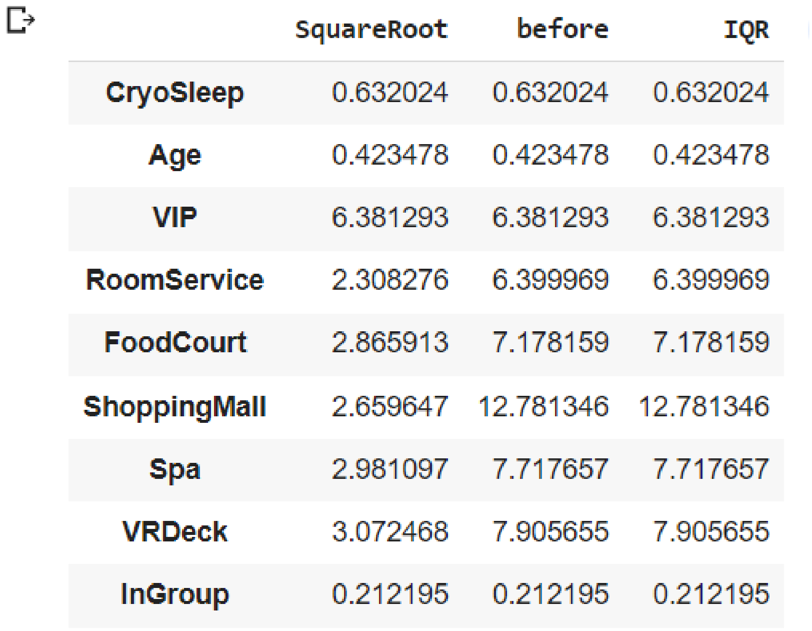


Removing Skewness:

As we can see the dataset is right-skewed and with zero values, neither log nor box-cox transformations can be used. Both of these transformations require that the data be positive and non-zero. In this case, the square root transformation and the Interquartile Range (IQR) method can be considered as alternatives.

Ultimately, the choice between the square root transformation and IQR scaling will depend on the specific characteristics of your data and the goals of your analysis. It's recommendable to test both methods and compare their performance before making a final decision.

Table showing the difference between the Square root transformation and IQR to normalize the data distibution



	SquareRoot	before	IQR
CryoSleep	0.632024	0.632024	0.632024
Age	0.423478	0.423478	0.423478
VIP	6.381293	6.381293	6.381293
RoomService	2.308276	6.399969	6.399969
FoodCourt	2.865913	7.178159	7.178159
ShoppingMall	2.659647	12.781346	12.781346
Spa	2.981097	7.717657	7.717657
VRDeck	3.072468	7.905655	7.905655
InGroup	0.212195	0.212195	0.212195

So here we select the square root transformation method. From the above table, we can see that the square root minimizes the highly right-skewed data from the dataset.

And we need to keep in mind that the outliers are not completely removed from the dataset, so while choosing the model for hyperparameter tuning, choose those models which are not sensitive towards the outliers. So

9. TOOLS USED

Python libraries make it easy for us to handle the data and perform typical and complex tasks with a single line of code.

Pandas – This library helps to load the data frame in a 2D array format and has multiple functions to perform analysis tasks in one go.

Numpy – Numpy arrays are very fast and can perform large computations in a very short time.

Matplotlib/Seaborn – This library is used to draw visualizations.

Sklearn – This module contains multiple libraries are having pre-implemented functions to perform tasks from data preprocessing to model development and evaluation.

XGBoost – This contains the eXtreme Gradient Boosting machine learning algorithm which is one of the algorithms which helps us to achieve high accuracy on predictions.

10.MATRICS OF CHOICE/KPI’s:

For this case study, we used Accuracy as KPI which is the Key Performance Indicator. The model will be evaluated using accuracy as the metric.

Accuracy is chosen as an evaluation metric for the titanic dataset is that the dataset is balanced, meaning that the number of survivors and non-survivors is roughly equal. This means that accuracy is a good measure of the model's performance, as it is not affected by the imbalance of the data.

It is also better to cross-verify the accuracy score with the other KPI like Confusion Matric, F1-score, Precision and Recall

Table showing the Tuned model scores

Sl.No	Tuned model name	Training Accuracy	Valuation Accuracy
1	Random Forest Classifier	86.12	79.06
2	Gradient Boosting Classifier	83.78	79.22
3	XG Boosting Classifier	81.13	78.71

Confusion Matric of Gradient Boosting classifier:

Predicted Label Actual label		Predicted Label		
		1	0	Total
Actual Label	1	970	319	1289
	0	223	1096	1319
	Total	1193	1415	2608

11.Prediction Result for Unseen Datapoints:

PassengerId	Transported
0013_01	TRUE
0018_01	FALSE
0019_01	TRUE
0021_01	TRUE
0023_01	TRUE
0027_01	FALSE
0029_01	TRUE
0032_01	TRUE
0032_02	TRUE

12.CONCLUSION:

We started with the data exploration where we got a feeling for the dataset, checked for missing data, and learned which features are important. During this process, we used seaborn and matplotlib to do the visualizations. During the data preprocessing part, we computed missing values, converted features into numeric ones, grouped values into categories, and created a few new features. Afterward, we started training 8 different machine learning models, and for each model, we did hyperparameter tuning and checked with the Accuracy score and also with a confusion matrix, classification report which consists of a precision, recall, and f-score picked one of them that is Gradient Boosting. But this model is overfitted so to improve or to avoid this we can use any one of the methods which are mentioned in the future scope of improvements.

13. FUTURE SCOPE OF IMPROVEMENTS:

As we can see the prediction of the data set of the finalized model, the model is overfitted that is the model is trained well but failed while predicting for the unseen data, so to avoid or to improve the prediction in future days we can use,

- Cross-validation
- Training with more data
- Removing features

- Early stopping the training
- Regularization
- Ensembling

14. REFERENCE:

https://betterprogramming.pub/titanic-survival-prediction-using-machine-learning-4c5ff1e3fa16

https://towardsdatascience.com/top-3-methods-for-handling-skewed-data-1334e0debf45

https://www.investopedia.com/terms/s/skewness.asp

https://colab.research.google.com/drive/1rd7_RNMxhGOdaOKvJxx1X8CkN-EewszX

https://www.freecodecamp.org/news/how-to-detect-outliers-in-machine-learning

https://www.geeksforgeeks.org/pandas-filling-nan-in-categorical-data/

https://blog.hubspot.com/marketing/choosing-kpis