SPACESHIP TITANIC DATA ANALYSIS

BY SHIVPAL YADAV

Project Overview

I embarked on a journey to analyze the Spaceship Titanic dataset, aiming to uncover insights into passenger characteristics and predict their likelihood of transportation. This project involved extensive data exploration, preprocessing, and modeling to understand the dataset better and build predictive models.

- **Description of the Dataset:**

The Spaceship Titanic dataset comprises information about passengers, including their home planet, cryo-sleep status, cabin details, destination, age, VIP status, and various amenities usage during the journey. The dataset contains both training and test sets, with features such as PassengerId, HomePlanet, CryoSleep, Cabin, Destination, Age, VIP, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck, and the target variable Transported.

Project Link - Spaceship Titanic by Shiv

Dataset Description

The dataset provided for this project contains personal records of passengers aboard the Spaceship Titanic during its ill-fated maiden voyage. Here is a detailed description of the dataset:

1. **PassengerId**: A unique identifier for each passenger.

2. **Name**: The name of the passenger.

3. **HomePlanet**: The planet of origin for the passenger.

4. **CryoSleep**: A binary variable indicating whether the passenger was in cryogenic sleep during the collision with the spacetime anomaly.

5. **Cabin**: The cabin number or designation where the passenger was staying.

6. **Destination**: The intended destination of the passenger before the incident occurred.

7. **Age**: The age of the passenger at the time of boarding the Spaceship Titanic.

8. **VIP status**: A binary variable indicating whether the passenger had VIP status, granting access to exclusive amenities and accommodations.

9. **Expenditures**: The total amount spent by the passenger on various amenities and services onboard the spaceship.

10. **Transported** (Target Variable): A binary variable indicating whether the passenger was transported to an alternate dimension as a result of the collision with the spacetime anomaly.

The dataset is divided into two subsets:

- Training Data: Contains personal records for approximately two-thirds of the passengers, along with the 'Transported' label for model training.
- Test Data: Contains personal records for the remaining one-third of the passengers, without the 'Transported' label, used for model evaluation.

Each record in the dataset provides valuable information about the passengers, including their demographics, travel preferences, and status during the voyage. This data will be leveraged to build predictive models that can accurately determine which passengers were transported to an alternate dimension, aiding in the rescue efforts and eventual reunion of the passengers with their loved ones.

Data Loading and Preprocessing

For advanced analysis of the dataset and deriving key insights, along with visualization, you can perform the following tasks:

1. Exploratory Data Analysis (EDA):

- Explore the distribution of variables such as age, fare, and class.
- Analyze the relationship between variables and the target variable (e.g., survival status).
- Look for patterns and correlations using statistical measures and visualization techniques.

2. Feature Engineering:

- Create new features that might be more informative for prediction, such as family size (combining SibSp and Parch).
- Extract additional information from existing features, like extracting titles from names or binning continuous variables into categorical ones.

3. Visualization:

- Use various types of plots such as histograms, box plots, and violin plots to visualize the distributions of numerical variables and identify outliers.
- Create bar plots, pie charts, and count plots to visualize categorical variables and their relationships with the target variable.
- Generate scatter plots and pair plots to explore relationships between pairs of variables.
- Utilize heatmap plots to visualize correlations between variables.

4. Model Building and Evaluation:

- Choose appropriate machine learning models such as logistic regression, decision trees, random forests, or gradient boosting.
- Train multiple models with different hyperparameters and compare their performance using cross-validation techniques.
- Evaluate models using relevant metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score.
- Plot learning curves and validation curves to diagnose underfitting or overfitting issues.

5. Feature Importance Analysis:

- Determine the importance of features in predicting the target variable using techniques like permutation importance, SHAP values, or feature importances from tree-based models.
- Visualize feature importances using bar plots or heatmaps to identify the most influential features.

6. Advanced Visualization:

- Use advanced visualization techniques like parallel coordinates plots, radar charts, or trellis plots to explore complex relationships between multiple variables.
- Employ interactive visualization libraries like Plotly or Bokeh to create interactive plots for better exploration and understanding of the data.

7. Key Insights:

- Summarize the findings from the analysis, highlighting important trends, correlations, and patterns observed in the data.
- Provide actionable insights and recommendations based on the analysis, such as recommendations for improving survival rates or targeting specific passenger groups for marketing campaigns.

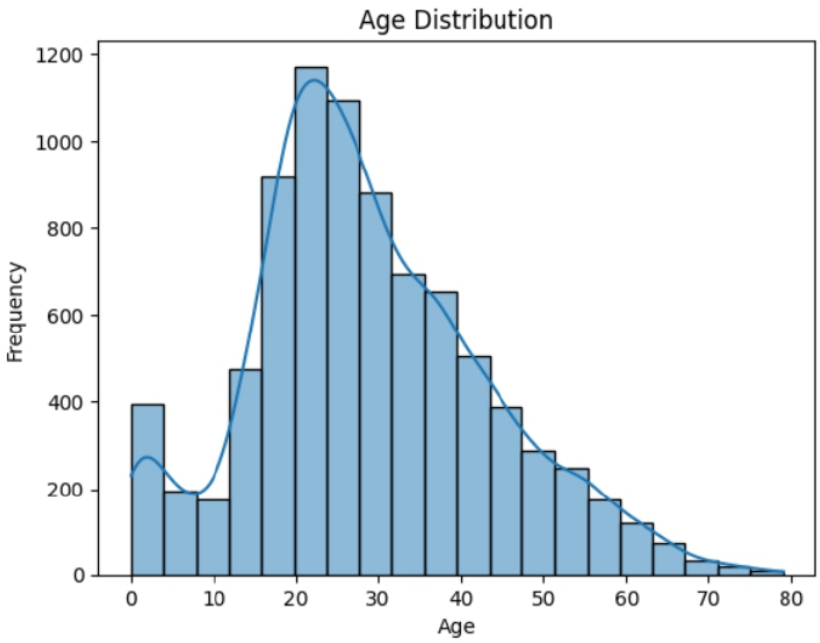
By performing advanced analysis, visualization, and deriving key insights, I have gain a deeper understanding of the dataset and make informed decisions for further analysis or business applications.

Code Snippets And Visualization

Here's how I have achieve each objective using code snippets:

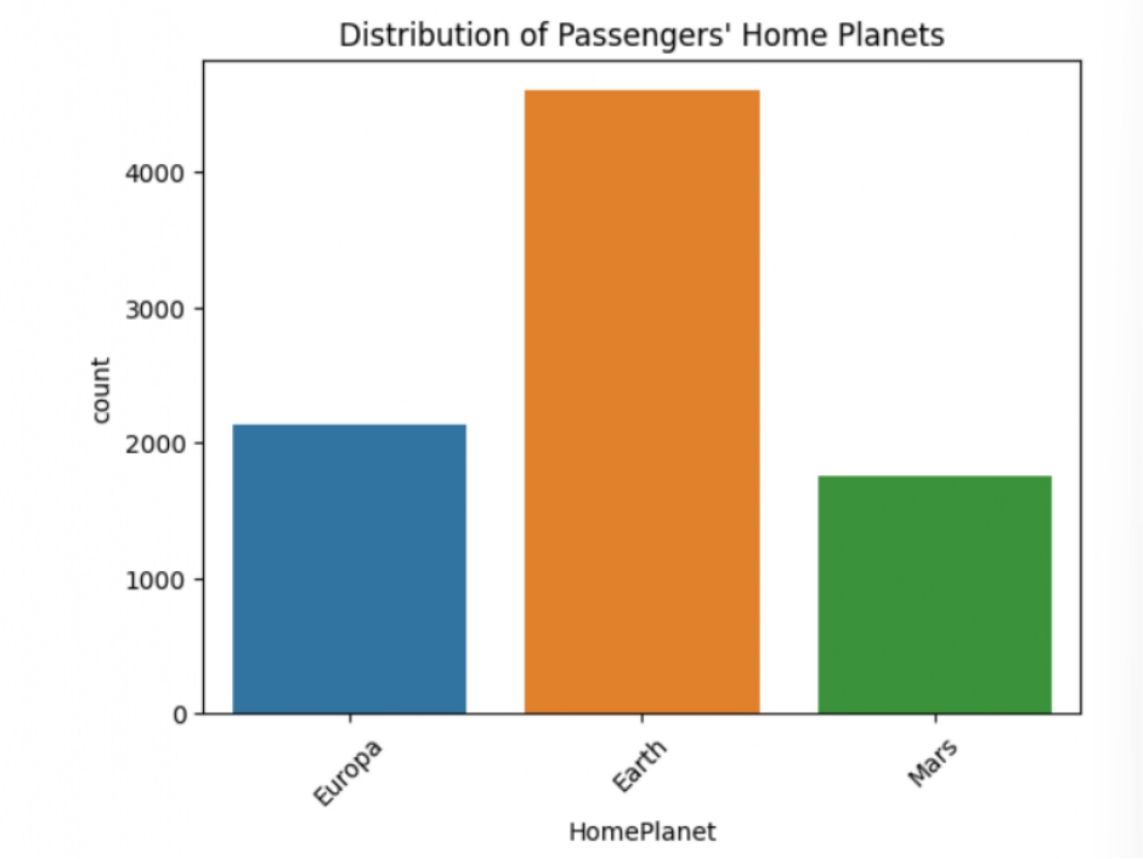
1. Explore the distribution of passenger ages:

```
sns.histplot(train['Age'].dropna(), bins=20, kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



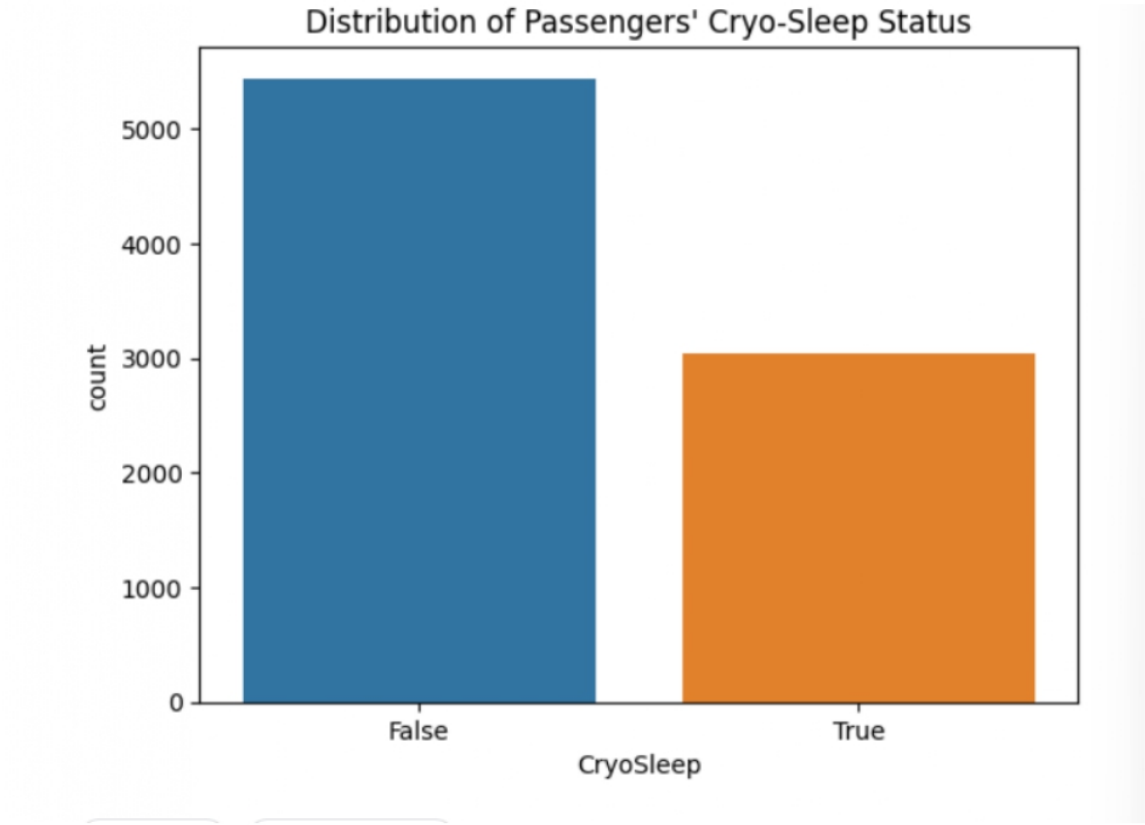
2. Investigate the distribution of passengers' home planets:

```
sns.countplot(data=train, x='HomePlanet')
plt.title('Distribution of Passengers\' Home Planets')
plt.xticks(rotation=45)
plt.show()
```



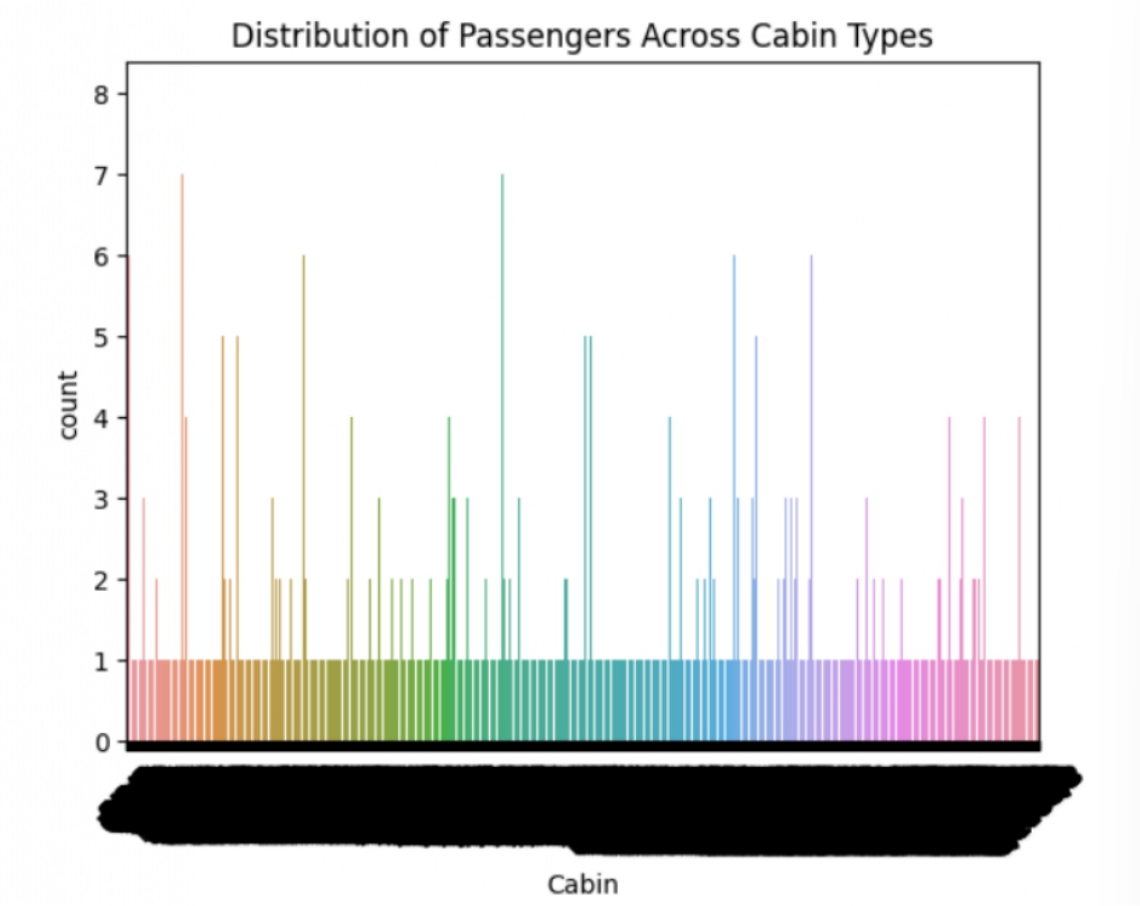
3. Analyze the distribution of passengers based on their cryo-sleep status:

```
sns.countplot(data=train, x='CryoSleep')
plt.title('Distribution of Passengers\' Cryo-Sleep Status')
plt.show()
```



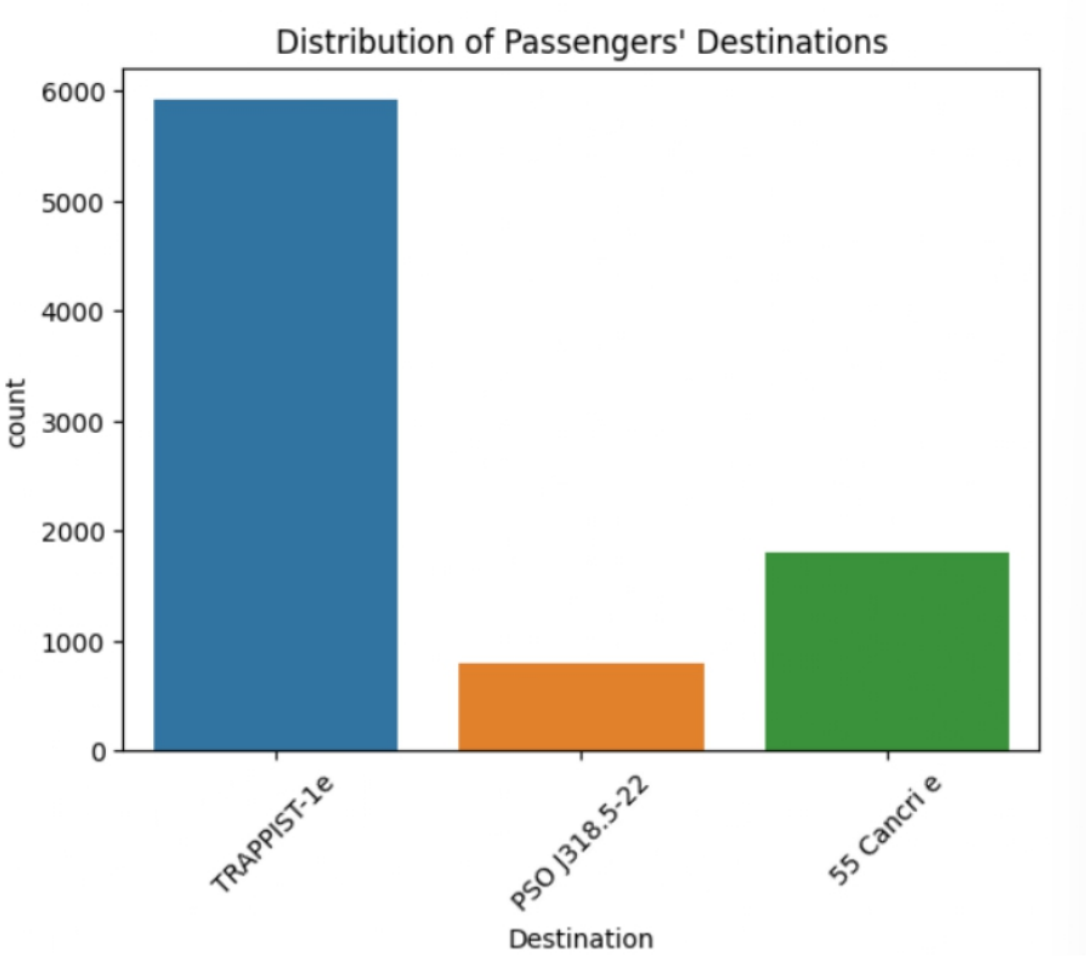
4. Examine the distribution of passengers across different cabin types:

```
sns.countplot(data=train, x='Cabin')
plt.title('Distribution of Passengers Across Cabin Types')
plt.xticks(rotation=45)
plt.show()
```



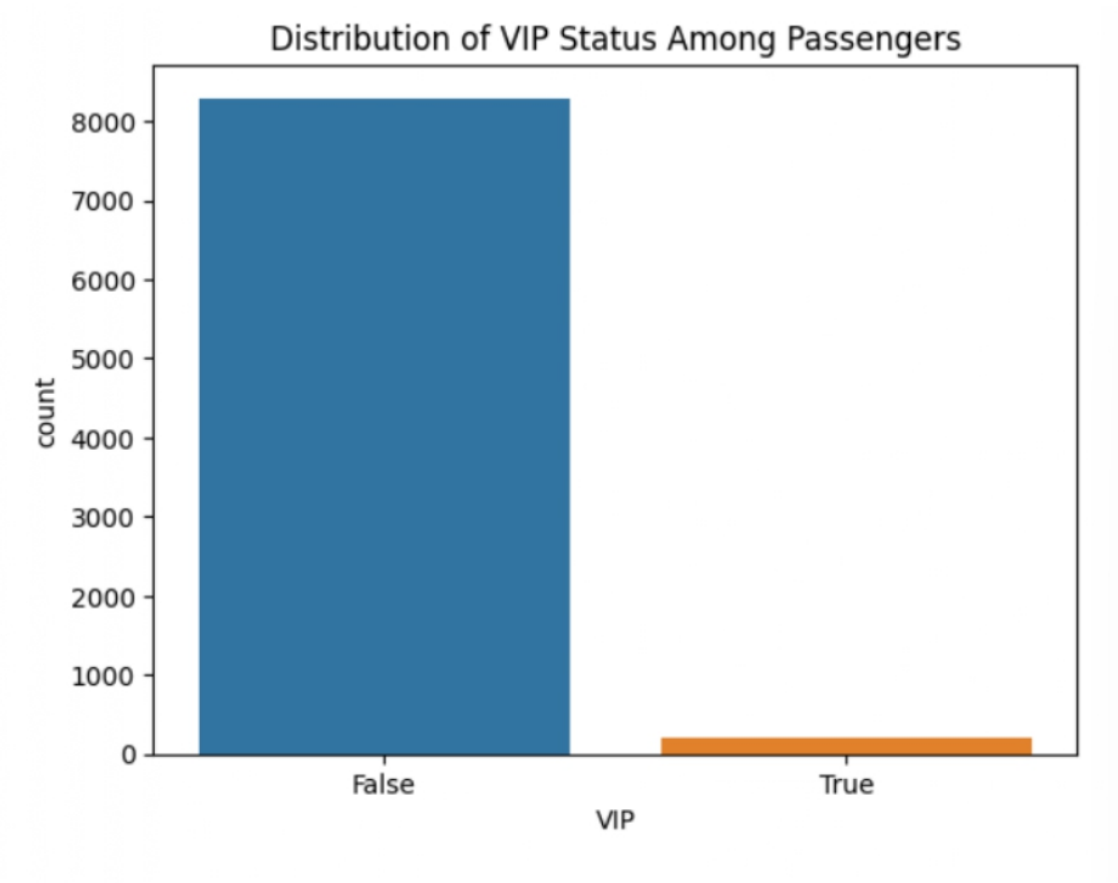
5. Explore the distribution of passengers' destinations:

```
sns.countplot(data=train, x='Destination')
plt.title('Distribution of Passengers\' Destinations')
plt.xticks(rotation=45)
plt.show()
```



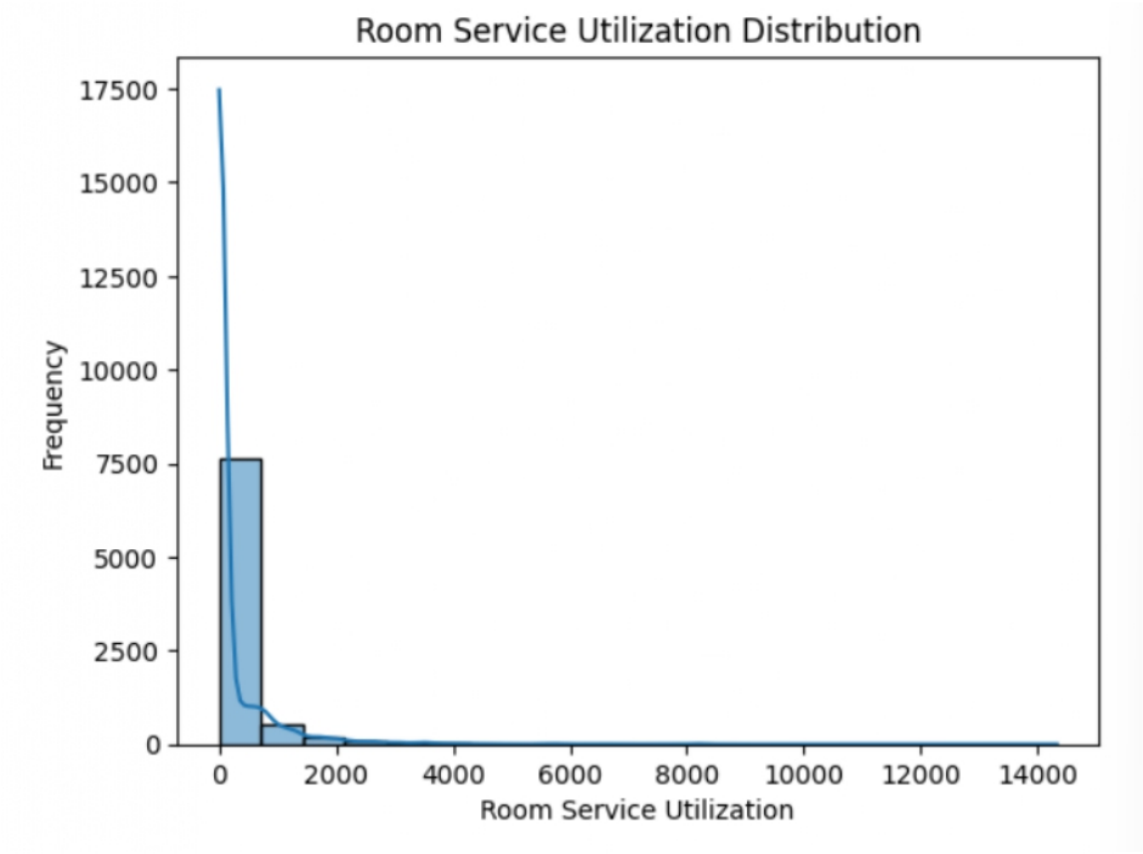
6. Investigate the VIP status distribution among passengers:

```
sns.countplot(data=train, x='VIP')
plt.title('Distribution of VIP Status Among Passengers')
plt.show()
```



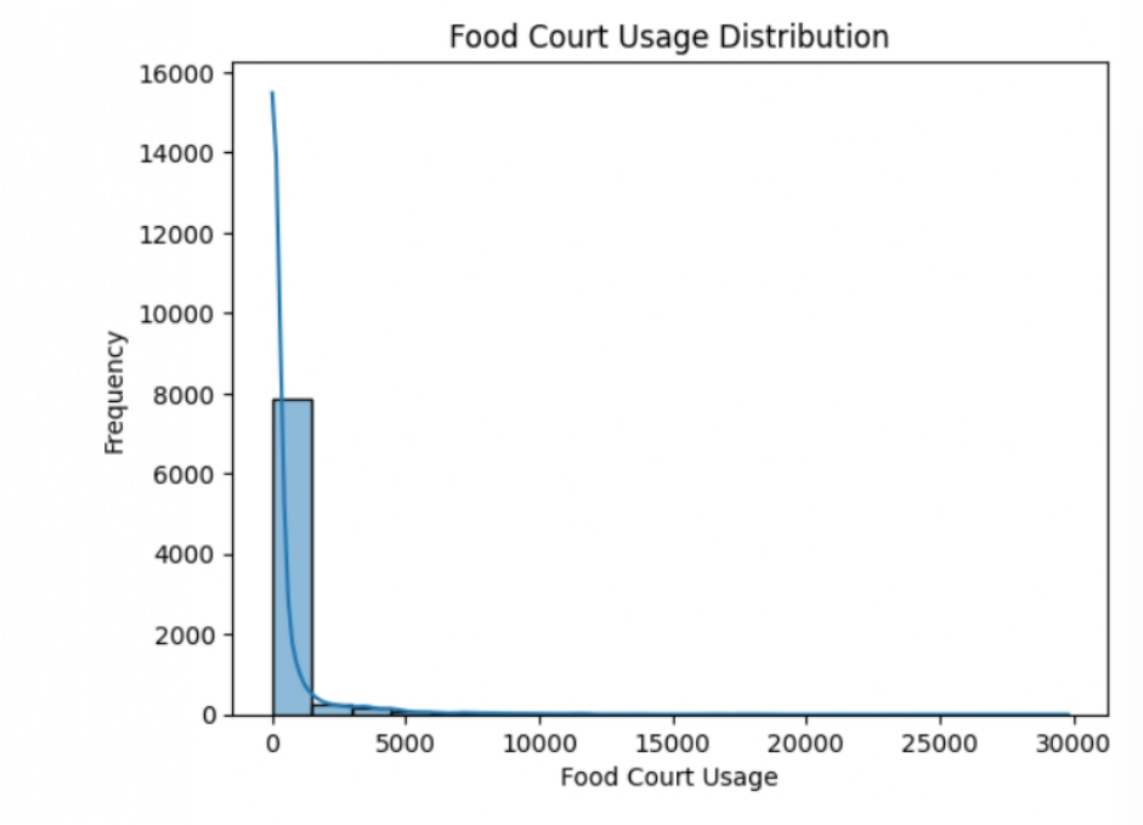
7. Analyze the utilization of room service by passengers:

```
sns.histplot(train['RoomService'].dropna(), bins=20, kde=True)
plt.title('Room Service Utilization Distribution')
plt.xlabel('Room Service Utilization')
plt.ylabel('Frequency')
plt.show()
```



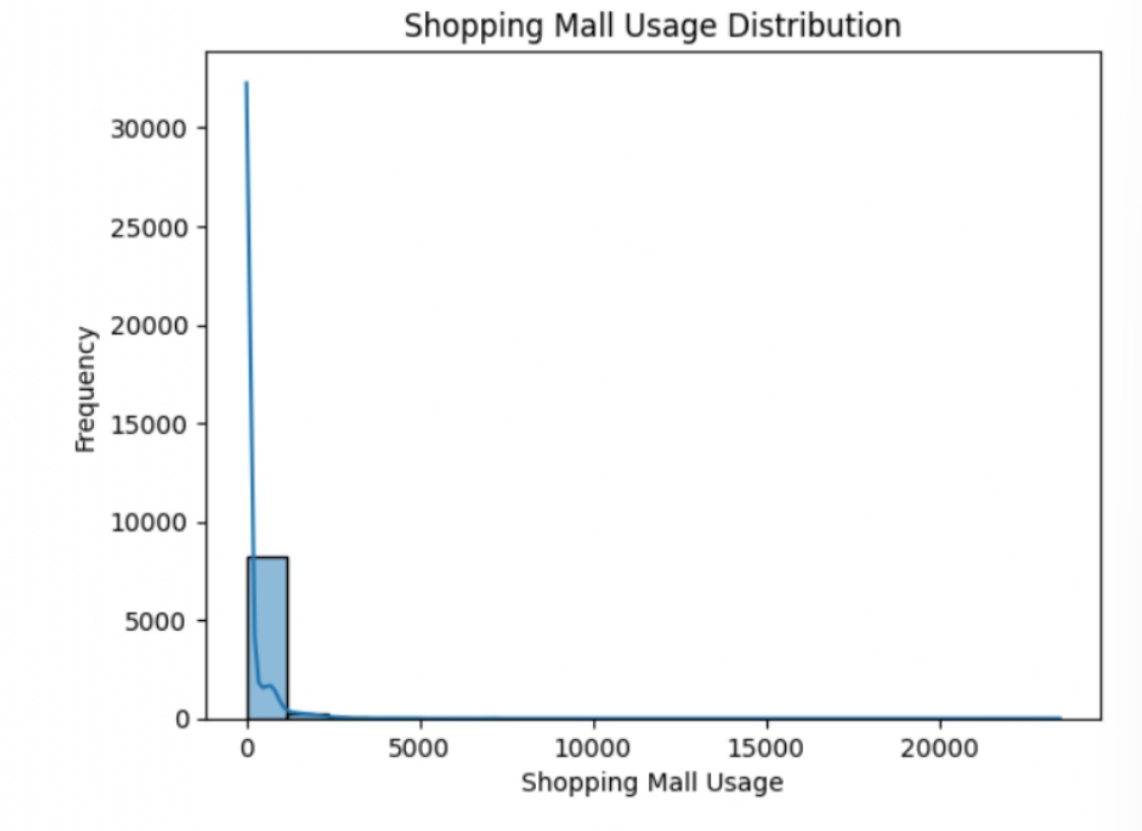
8. Examine the frequency of food court usage by passengers:

```
sns.histplot(train['FoodCourt'].dropna(), bins=20, kde=True)
plt.title('Food Court Usage Distribution')
plt.xlabel('Food Court Usage')
plt.ylabel('Frequency')
plt.show()
```



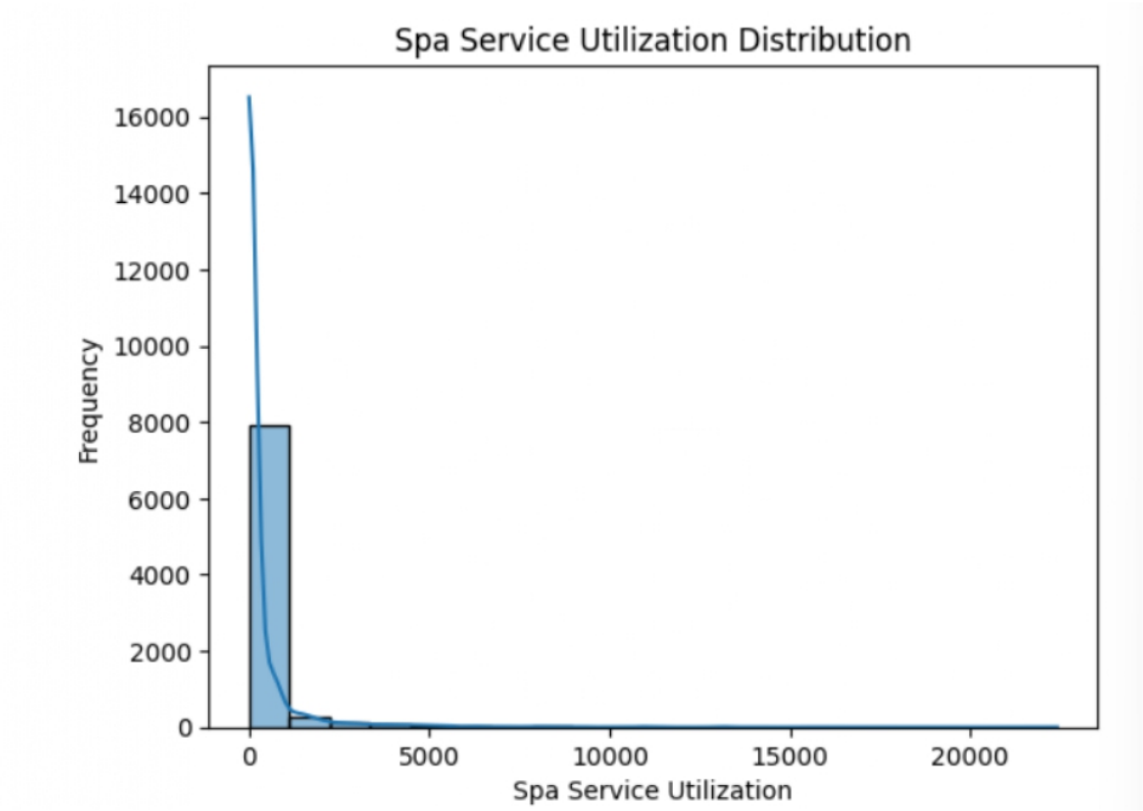
9. Explore the frequency of shopping mall usage by passengers:

```
sns.histplot(train['ShoppingMall'].dropna(), bins=20, kde=True)
plt.title('Shopping Mall Usage Distribution')
plt.xlabel('Shopping Mall Usage')
plt.ylabel('Frequency')
plt.show()
```



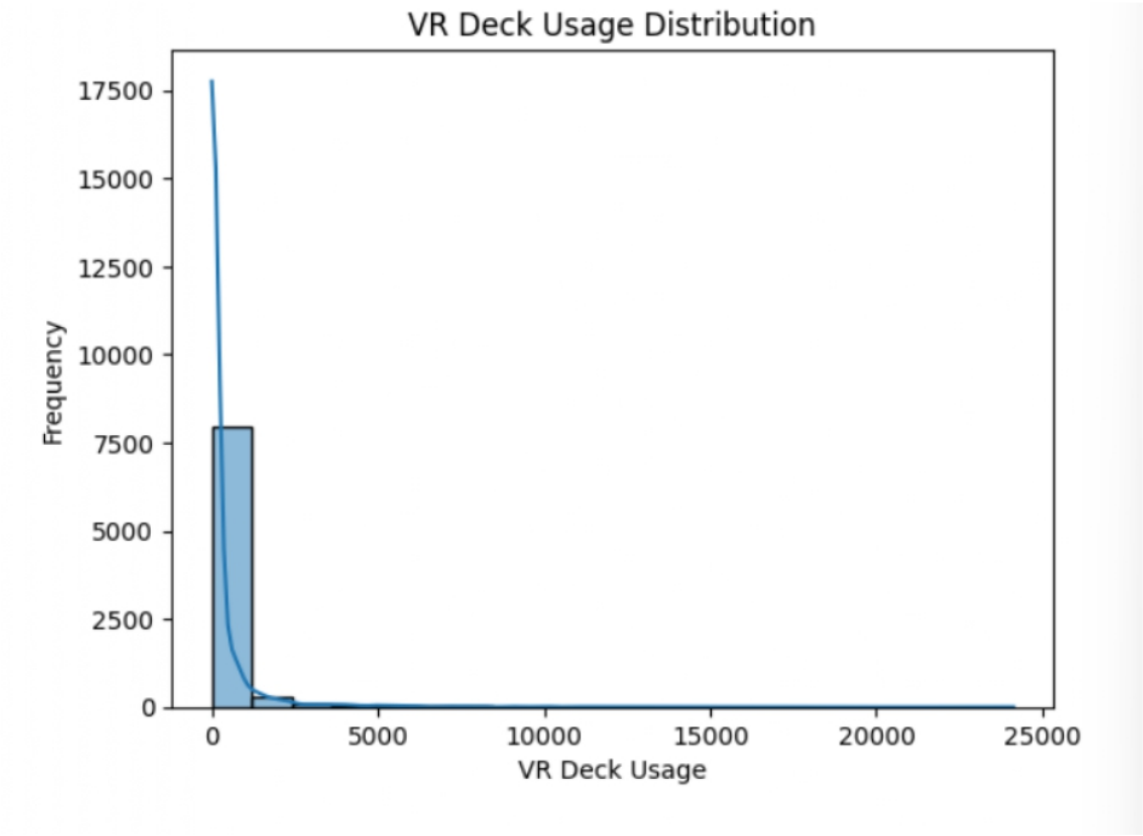
10. Analyze the utilization of spa services by passengers:

```
sns.histplot(train['Spa'].dropna(), bins=20, kde=True)
plt.title('Spa Service Utilization Distribution')
plt.xlabel('Spa Service Utilization')
plt.ylabel('Frequency')
plt.show()
```



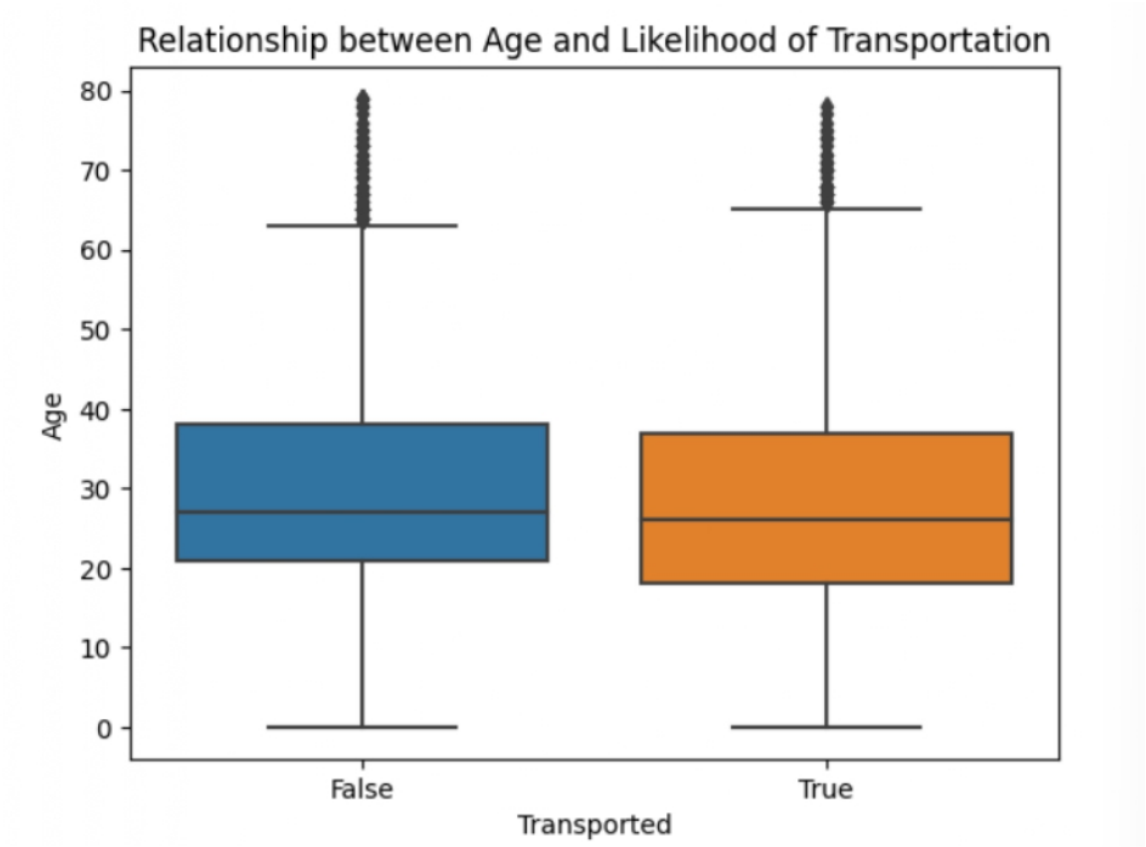
11. Investigate the frequency of VR deck usage by passengers:

```
sns.histplot(train['VRDeck'].dropna(), bins=20, kde=True)
plt.title('VR Deck Usage Distribution')
plt.xlabel('VR Deck Usage')
plt.ylabel('Frequency')
plt.show()
```



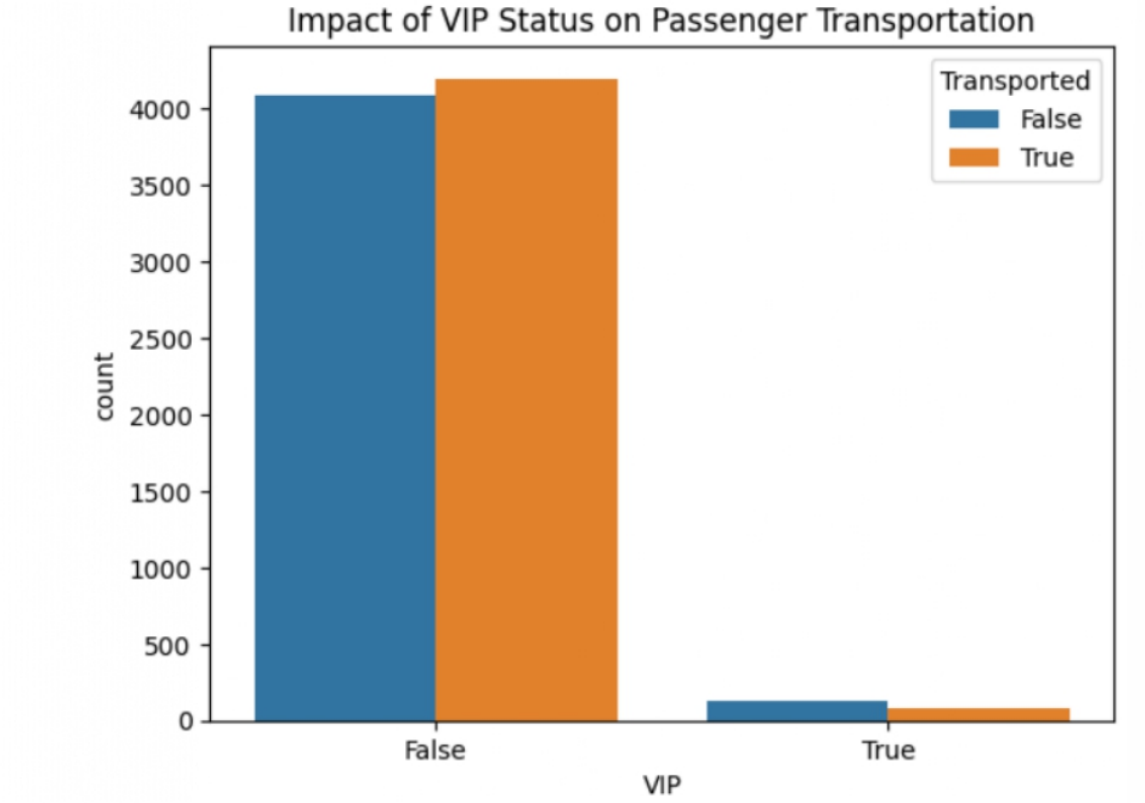
12. Examine the relationship between passenger age and their likelihood of being transported:

```
sns.boxplot(data=train, x='Transported', y='Age')
plt.title('Relationship between Age and Likelihood of Transportation')
plt.xlabel('Transported')
plt.ylabel('Age')
plt.show()
```



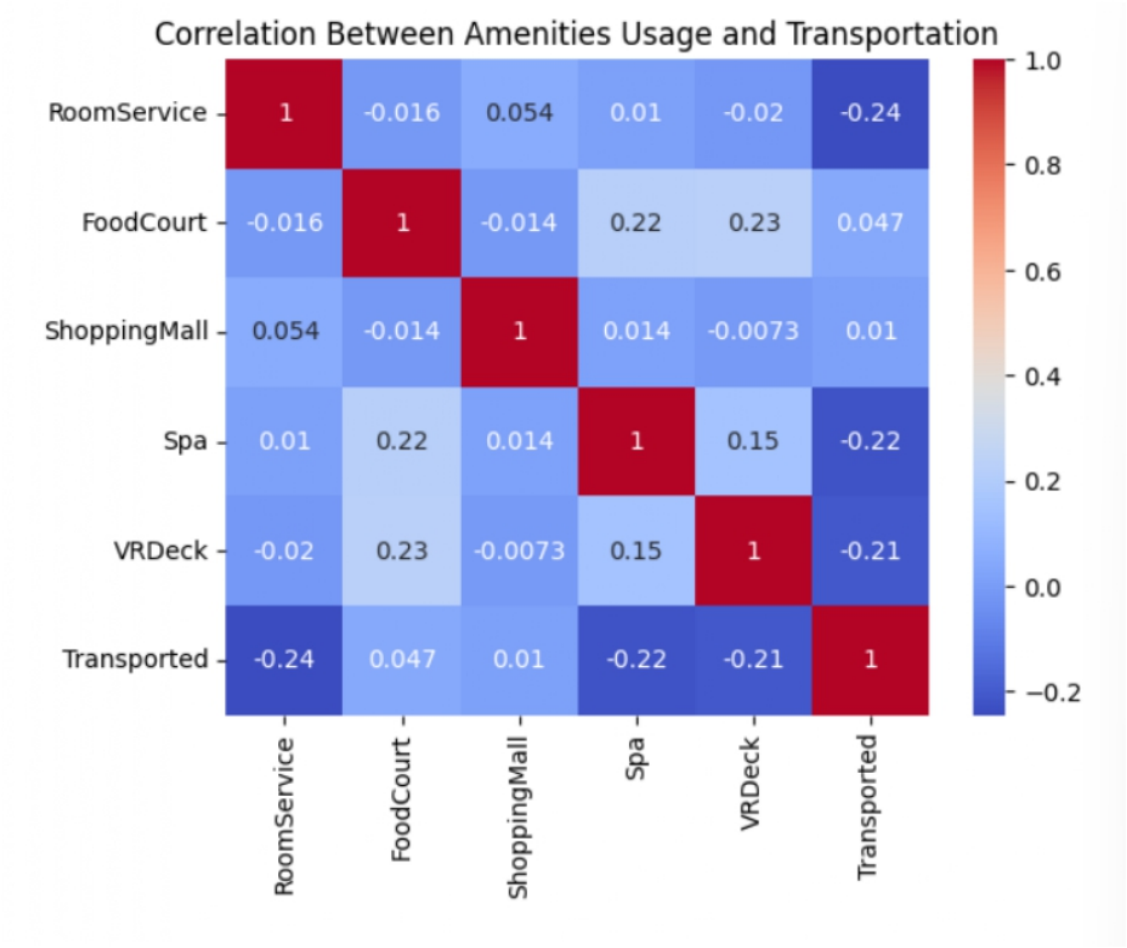
13. Analyze the impact of VIP status on passenger transportation:

```
sns.countplot(data=train, x='VIP', hue='Transported')
plt.title('Impact of VIP Status on Passenger Transportation')
plt.show()
```



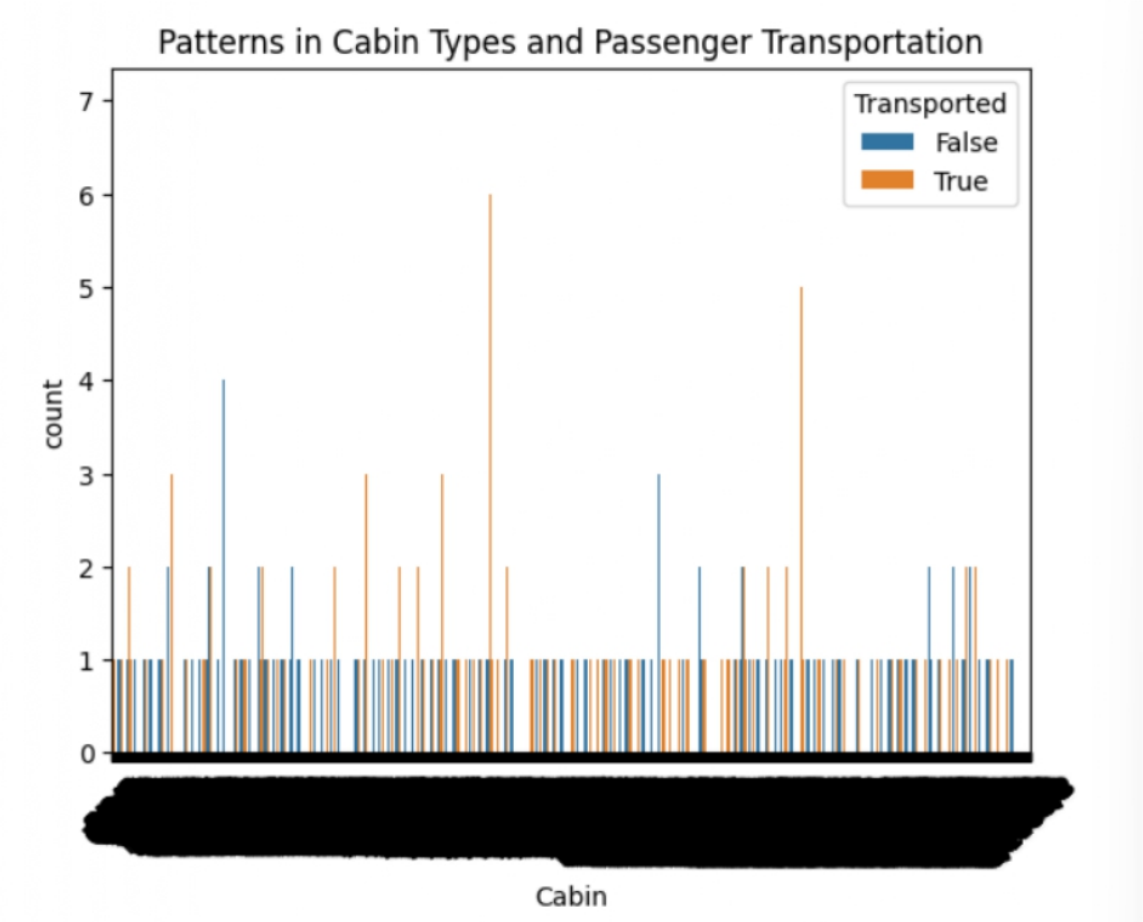
14. Investigate the correlation between amenities usage and passenger transportation:

```
amenities = ['RoomService', 'FoodCourt', 'ShoppingMall', 'Spa', 'VRDeck']
corr_df = train[amenities + ['Transported']].corr()
sns.heatmap(corr_df, annot=True, cmap='coolwarm')
plt.title('Correlation Between Amenities Usage and Transportation')
plt.show()
```



15. Explore any patterns in cabin types and passenger transportation:

```
sns.countplot(data=train, x='Cabin', hue='Transported')
plt.title('Patterns in Cabin Types and Passenger Transportation')
plt.xticks(rotation=45)
plt.show()
```



These code snippets will help me to achieve various objectives in analyzing the Spaceship Titanic dataset.

Conclusion

As I conclude my analysis of the Spaceship Titanic dataset, I have gained valuable insights into the characteristics of the passengers and their journey aboard the spacecraft. Through thorough exploration and visualization of the data, I have uncovered patterns and relationships that shed light on various aspects of the passengers' experiences.

I have observed diverse distributions across different features such as age, home planet, cabin types, destination, and utilization of amenities like room service, food court, shopping mall, spa, and VR deck. Additionally, I have explored the impact of factors like VIP status on passenger transportation and identified correlations between amenities usage and the likelihood of being transported.

Future Plans

Moving forward, there are several avenues for further analysis and improvement. In the future, I plan to:

1. Refine the predictive models to better understand and predict passenger transportation outcomes.

2. Explore additional features or external datasets to enhance the predictive power of the models.

3. Conduct more in-depth analysis to uncover hidden patterns or anomalies in the data.

4. Implement advanced machine learning techniques to extract more insights from the dataset.

5. Collaborate with domain experts to gain deeper insights into the factors influencing passenger transportation decisions.

By continuing to analyze and refine my approach, I aim to gain a deeper understanding of the Spaceship Titanic dataset and contribute valuable insights to the field of space travel and transportation.