Titanic Spaceship

The dataset consists of personal records for (8693) of the passengers.

The 15 main features and their descriptions are listed below:

* PassengerId - A unique identifier for each passenger in the format 'gggg\_pp', where 'gggg' represents the group and 'pp' indicates the individual's number within that group.
* HomePlanet - The planet from which the passenger departed.
* CryoSleep - Indicates whether the passenger was in suspended animation during the voyage. Passengers in cryosleep were confined to their cabins.
* Cabin - The passenger's cabin number, which consists of the deck, room number, and side (port or starboard).
* Destination - Indicates the destination planet to which the passenger was travelling: TRAPPIST-1e, 55 Cancri e, or PSO J318.5-22.
* Age - The passenger's age.
* VIP - Indicates whether the passenger had paid for VIP services.
* RoomService / FoodCourt / ShoppingMall / Spa / VRDeck - The amount the passenger was billed for using each of Spaceship Titanic's luxury amenities.
* Name - First and last names of passengers.
* Transported - Indicates whether the passenger was transported to another dimension during the collision.

I did some analysis on the dataset using R language.

**HomePlanet:**

A graph of a graph

Description automatically generated with medium confidence

Homeplanet seems to be correlated with Transported.

Notably, the distribution of Transported values in the NA data is much closer to data labeled Mars compared to Earth.

So, I replace the Na value with Earth.

**Cryosleep:**

A graph of a bar chart

Description automatically generated with medium confidence

From the graph it seems that CryoSleep column is strongly correlated with Transported. I'd like my chances of staying on the ship more if I was conscious when the crash happened.

A graph with different colored squares

Description automatically generated

Destination looks to be another important column. The proportion of passengers Transported differs significantly depending on their Destination.

**VIP by category:**

A graph with a red and blue square

Description automatically generated

VIP doesn't seem to be too useful due to the similar proportions of Transported and the small number of TRUE observations.

Total Spent:

Here I counted the total money spent on RoomService , FoodCourt , ShoppingMall ,Spa, VRDeck.

A graph of a bar chart

Description automatically generated with medium confidence

Spending on amenities is correlated with Transported.

**ShoppingMall , Roomservice , FoodCourt , Spa, VRDeck :**

**A diagram of distribution of data

Description automatically generated with medium confidence**

They aren't normal distributions, so the values of NA are considered as the median.

**Age by category:**

**A graph of red and grey bars

Description automatically generated**

There's multiple features that can be extracted from the Cabin and PassengerId columns. From looking at the data, you can see that the last part of the PassengerId column represents the number of the person in their group. I extracted that data to create a group size column called size. Because there are such small amounts of large groups, I decided to count groups of four or larger simply as four.

**Plot the group size feature:**

**A graph with green and red squares

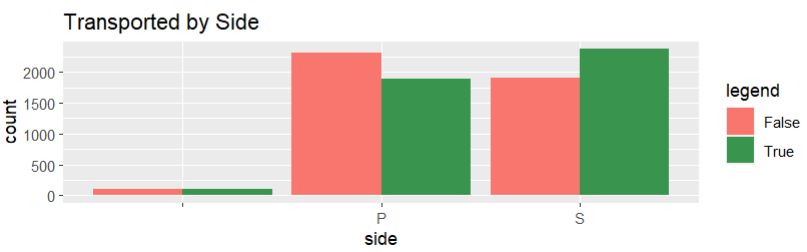
Description automatically generated**

From the above graph we conclude that singles are more likely to stay in the Ship.

**Plot Cabin:**

**A graph with green and red bars

Description automatically generated**



**Deck** - There are some interesting patterns between the different decks. Passengers on decks B and C were significantly more likely to be transported, while passengers on deck E were less likely to be transported.

**Side** - There appears to be a noticeable difference in the likelihood of being transported based on which side of the ship passengers are on. Passengers on the port side were safer from being transported than those on the starboard side.

**I split age into age groups for analysis purposes.**

**A graph with green and red bars

Description automatically generated**

Based on the above chart, it appears that passengers under the age of 16 were much more likely to be transported than other age groups. For the other age groups, there is a relatively balanced distribution between those who were taken and those who were not taken.

**Correlation:**

A graph with numbers and letters

Description automatically generated with medium confidence

There is low correlation between these attributes.