583 Exploratory Data Analysis

Henry Luo, Renghe Tang, Weijia Lyu

2023-03-10

The current project attempts to investigate if there are certain features that make individual passengers aboard the fictional spaceship Titanic more likely to be transported to the alternative dimension. If there are features that make passengers more vulnerable to the space-time anomaly, understanding their relationship with the probability of getting transported could potentially help protect passengers on other ships in the surrounding area.

The "Space Titanic" Kaggle competition (https://www.kaggle.com/competitions/spaceship-titanic/data) splits the Space Titanic passenger records into a training set and a test set for the sake of machine learning. Since the focus of the current project is to understand the relationship between different passenger features and the probability of being transported, the scope of the exploratory analysis is limited to the training set since the test set is missing the response variable of whether the passengers are transported.

Summary of Variables

The dataset has 8693 passenger records with 14 variables. 8 variables are of the type character. Variables `HomePlanet`, `CryoSleep`, `Destination`, `VIP`, and `Transported` appear to be encoding categorical information. They are changed to the type factor in this analysis. 6 of the variables are numeric, encoded as doubles.

According to the variable descriptions on the Kaggle competition website, `PassengerId` and `Cabin` contain multiple levels of information. `PassengerId` consists of the group number and passenger's Id in that group. `Cabin` consists of deck number, room number, and which side of the ship the room is at (port side or starboard side). They can be separated in different categorical variables. `PassengerId` is separated to `Group` and `InGroupId`. `Cabin` is separated to `Deck`, `RoomNum`, and `Side`. The number of variables is thus expanded to 17, including 8 factor variables, 8 numeric variables, and passenger names (`name`) encoded as character.

Explore Categorical Variables

Figure 1 outlines the first four categorical variables and the number of passengers who got transported compared to those not transported within each of their categories. The upper left graph indicates that among the passengers who chose to remain in cryosleep, much more of them were transported compared to those not transported, while a lot more passengers who chose to be awake were not transported than those who were awake and transported. It suggests that people in cryosleep are more likely to be transported. The graph on the upper right side suggests that there seem to be no considerable differences between transported and not transported across VIPs and non-VIPs. The lower left graph demonstrates that passengers from Europa and perhaps Mars were more likely to be transported than not, while passengers from Earth were less likely to be transported. The graph on the lower right side suggests that passengers going to 55 Cancri e were more likely to be transported than not, while those going to TRAPPIST-1e were less likely to be transported. The NA data, missing records, across all four variables show no considerable difference in the chance of being

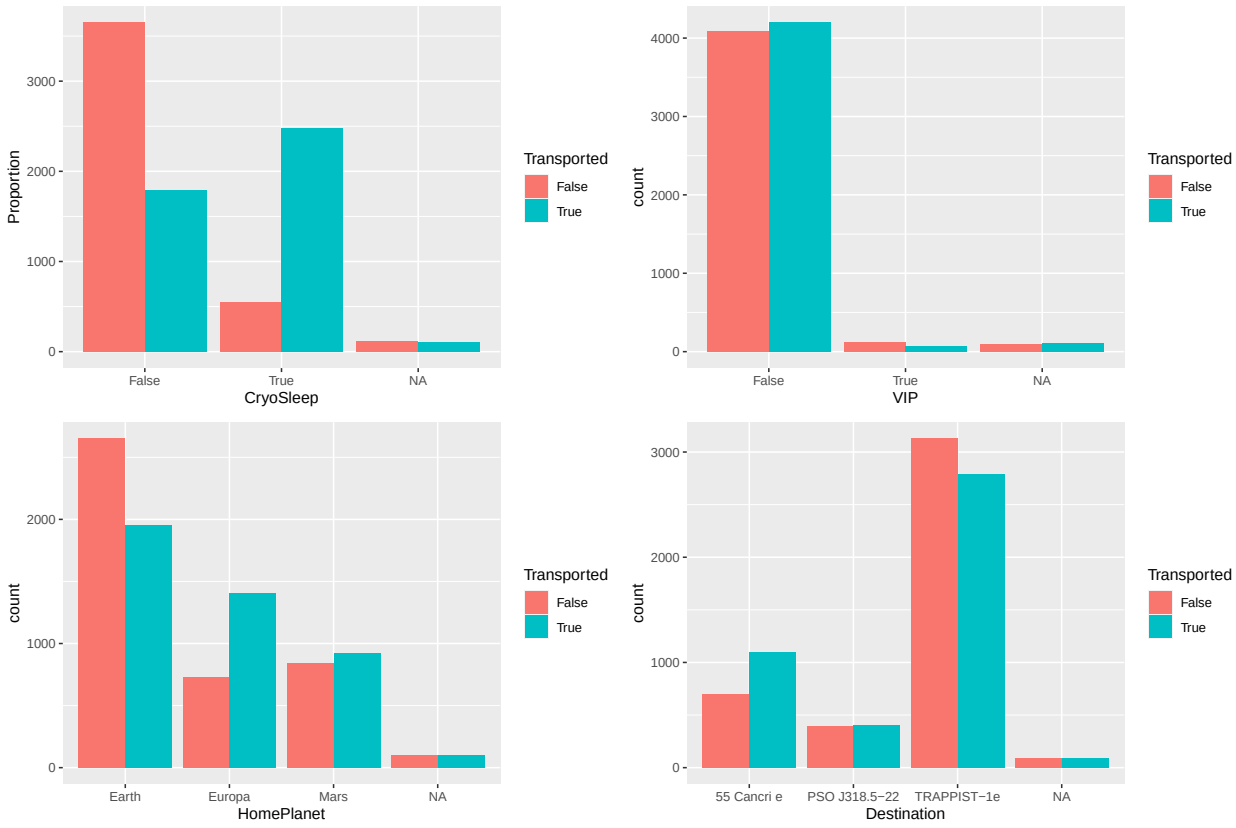

Figure 1: Categorical variables by Transported

transported. In conclusion, Figure 1 suggests that `Cryosleep`, `HomePlanet`, and `Destination` seem to affect a passenger's chance of being transported.

Figure 2 demonstrates the number of people being transported compared to those not transported within each category of the additional categorical variables from expanding `PassangerId` and `Cabin`. The upper left graph shows that passengers' chance of being transported seems to differ by the deck they are on, with being on decks B,C, and G having a higher chance of being transported than on other decks. The upper right graph shows that passengers on the port side were less likely to get transported, while those on the starboard side were more likely to be transported than not. Finally, the lower left graph suggests that the chance of being transported also differs by passenger's id in their travel group. Passengers with mission data again show no difference in the chance of being transported or not. In conclusion, `Deck`, `Side`, and `InGroupId` all seem to affect a passenger's chance of being transported.

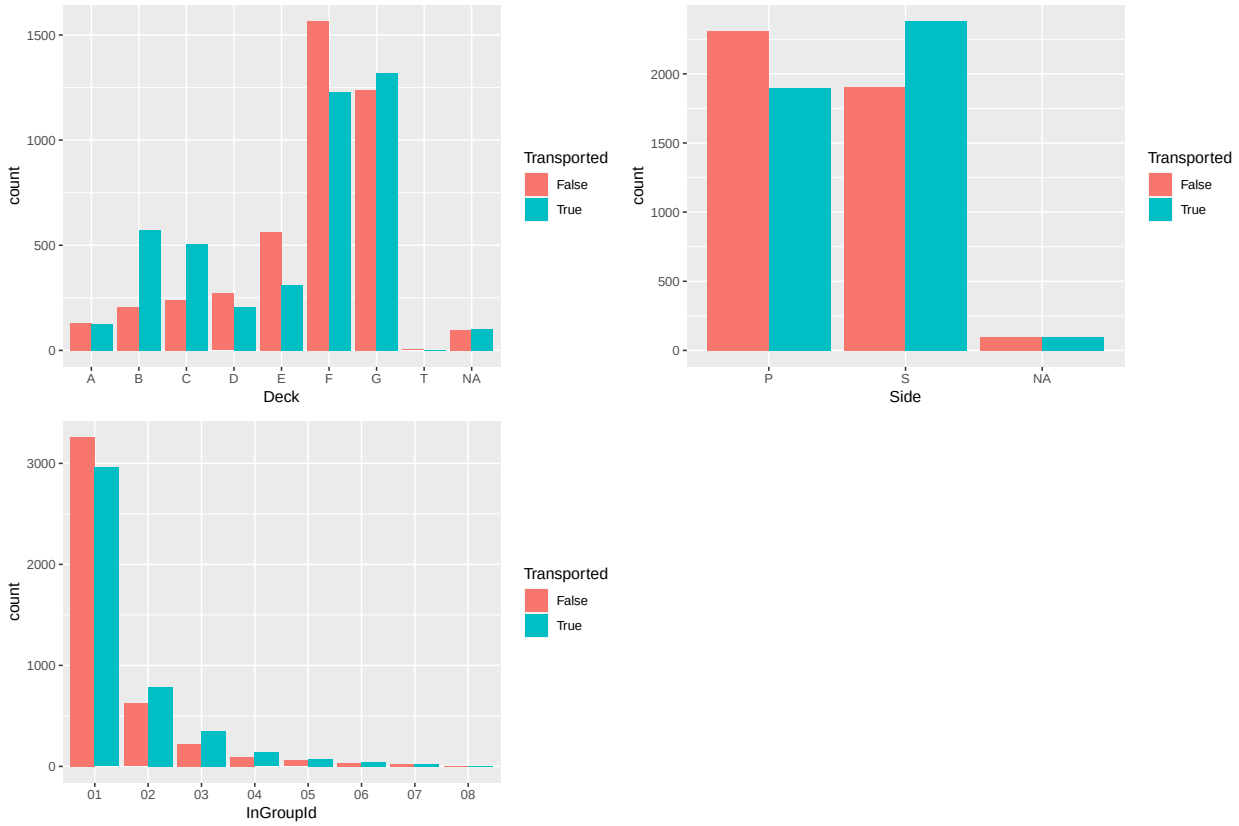

Figure 2: Expanded Categorical variables by Transported

Explore Numeric Variables

Table 1 contains the summary statistics of the important numerical variables. The majority of them record passengers' spending on the ship's recreation facilities, which contain many zero values since more than half of the passengers chose to remain in cryosleep (to preserve the body over the prolonged voyage). No unit was provided with the variables.

Table 1: Summary Statitics of Numeric Variables

	mean	sd	median	min	max	range
Age	28.82793	14.48902	27	0	79	79
RoomService	224.68762	666.71766	0	0	14327	14327
FoodCourt	458.07720	1611.48924	0	0	29813	29813
ShoppingMall	173.72917	604.69646	0	0	23492	23492
Spa	311.13878	1136.70553	0	0	22408	22408
VRDeck	304.85479	1145.71719	0	0	24133	24133

Figure 3 shows the distribution of the important numerical variables. The distribution of passenger age is relatively skewed. It seems to be heavily affected by a large number of 0-year-old infants on board, and the number of children and teenagers is much less than that of adults. The other variables reflecting passengers' spending are all heavily skewed. Most of the passengers only spend a small sum of money on recreation, while some rich passengers spend incredibly more.

Further investigation is conducted on `Age`. Figure 4 shows the age distribution of passengers who were transported compared with those who were not. It appears that children and teenagers before 20 years of age were more likely to be transported, especially younger children and infants. Much more 0-year-old infants were transported than not.

The relationships between the numerical variables and the transportation outcome were investigated in Figure 5. Records belonging to Transported passengers were marked in red. It seems that several variables, such as `RoomService`, `Spa`, and `VRDeck` are affecting the chance of being transported. For example, the more one spends on `Spa` and `VRDeck`, the less likely one is transported.

Since the numerical data are highly skewed towards zero, it is difficult to check for collinearity between variables from Figure 5. Therefore, a correlation matrix was calculated to determine if there is severe multi-collinearity between the numeric variables.

Exploratory Analysis

A preliminary analysis using random forest was conducted to investigate the importance of different variables in predicting the if a passenger is transported. Random forest is a good choice since it has fewer assumptions and can handle outliers. The model achieved an OOB estimate of error rate of around 0.2, indicating that the model fit is decent for the purpose of this preliminary analysis. Figure 6 outlines the result of the variable importance obtained from the random forest model. Confirming the findings from previous sections, `CryoSlee`, `Spa`, and `VRDeck` are among the high-importance variables, suggesting that they indeed play crucial roles in affecting a passenger's chance of being transported.

Planning for Logistic Regression Analysis

The goal of the current project is to understand the relationship between passenger features and the probability of them being transported. It seems that logistic regression would provide a better level of interpretability that is more suitable for the purpose of the current investigation. It is easier to see the direction of each variable's influence on the probability of being transported and compute the amount of that influence. It would answer the questions of how likely a person with specific features would be transported and how much safer people could become if they change certain features. It could also help devise emergency plans such as prioritizing evacuating infants from Europa on the ship's starboard side since they are more vulnerable to

the space-time anomaly. The data is suitable for logistic regression since the response is binary, and logistic regression can handle the mixture of numeric and categorical data.

An important assumption of logistic regression is that there is no multi-collinearity between the variables. Since the numerical data are highly skewed towards zero, it is difficult to check for collinearity between variables from Figure 5. Therefore, a correlation matrix was calculated to determine if there is severe multi-collinearity between the numeric variables and presented in Table 2.

Table 2: Correlation Matrix of Numeric Variables

	Age	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck
Age	1.0000000	0.0687230	0.1304210	0.0331326	0.1239703	0.1010073
RoomService	0.0687230	1.0000000	-0.0158889	0.0544803	0.0100795	-0.0195815
FoodCourt	0.1304210	-0.0158889	1.0000000	-0.0142277	0.2218905	0.2279954
ShoppingMall	0.0331326	0.0544803	-0.0142277	1.0000000	0.0138786	-0.0073217
Spa	0.1239703	0.0100795	0.2218905	0.0138786	1.0000000	0.1538212
VRDeck	0.1010073	-0.0195815	0.2279954	-0.0073217	0.1538212	1.0000000

There does not seem to be too much multi-collinearity between the numerical variables, but it should be cautioned when fitting the logistic model as some variables have correlations of around 0.2.

Although logistic regression does not assume that its numeric predictor variables follow any specific distribution, the high level of skewness of the variables in this data set could still have effects on the logistic model. The skewed variables could be log-transformed to see if it improves the model fit. Likelihood Ratio Test could be used to inform if removing extremely skewed variables improves model fit.

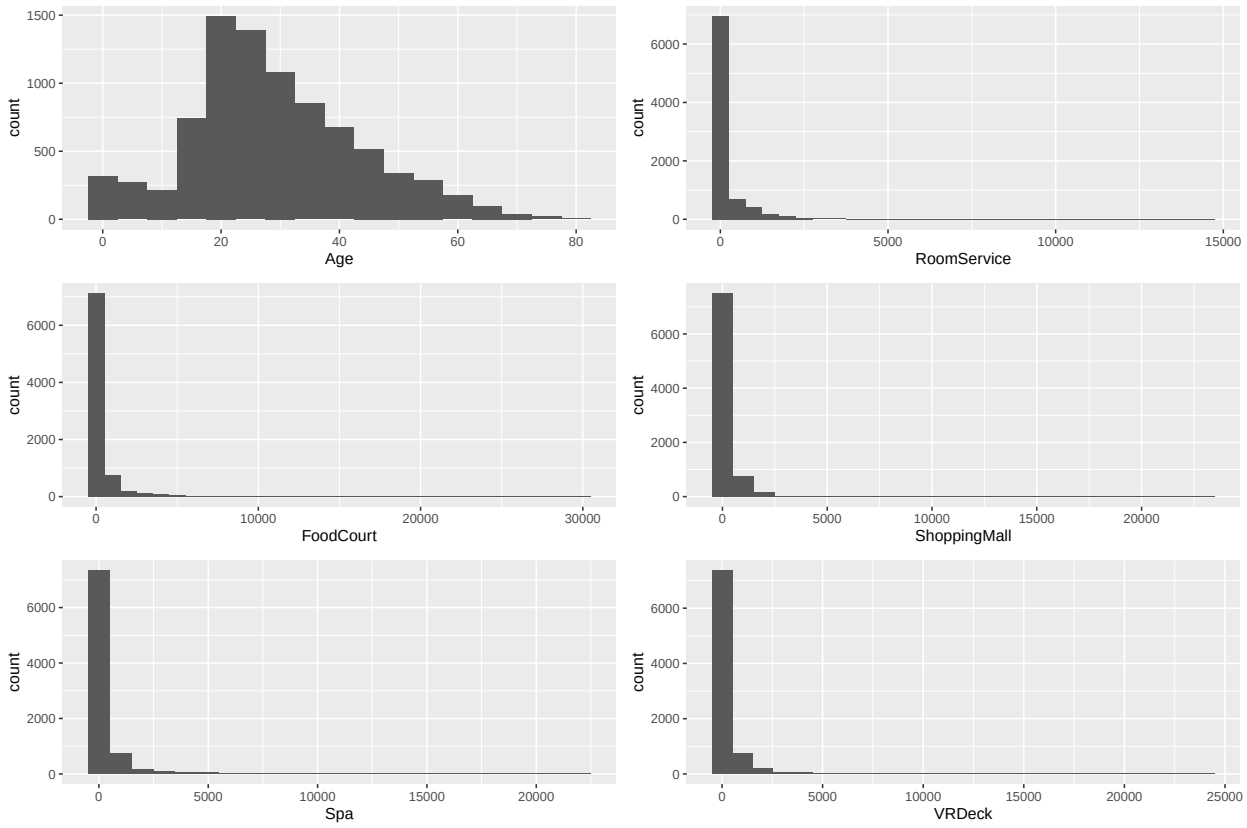

Figure 3: Distribution of Numeric Variables

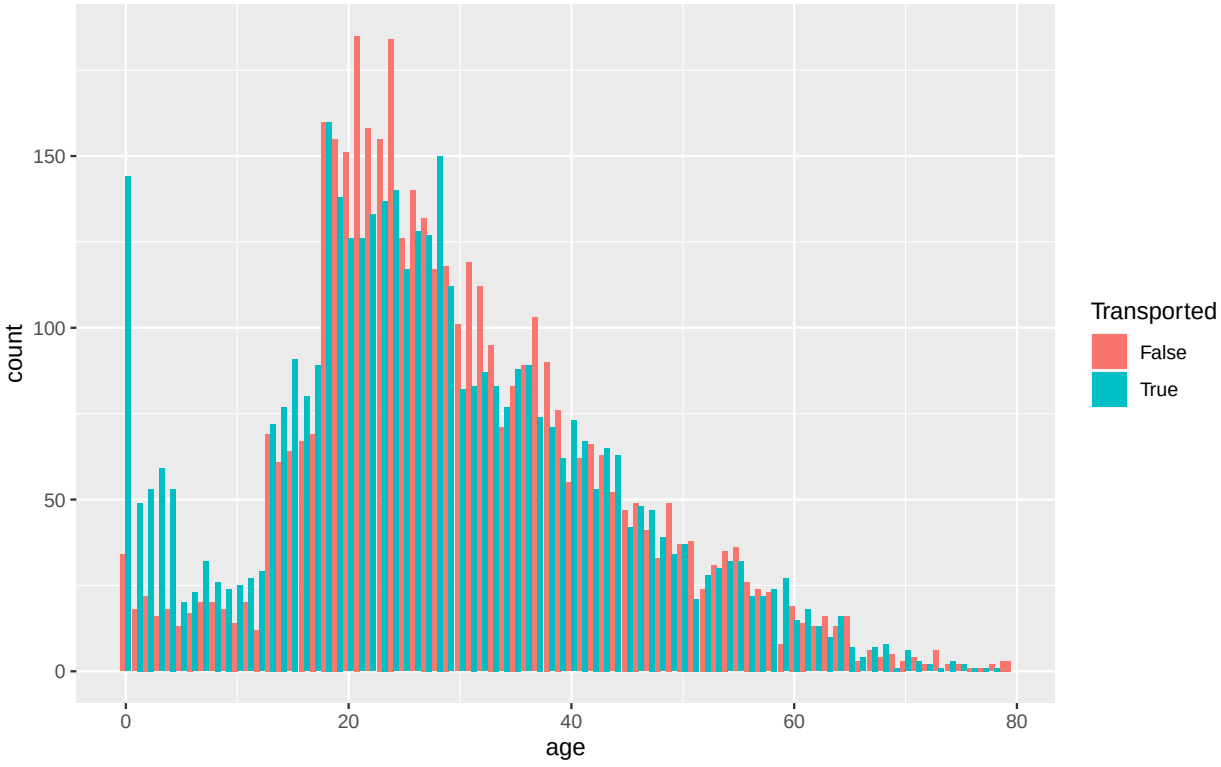

Figure 4: Age of Transported Versus Not Transported

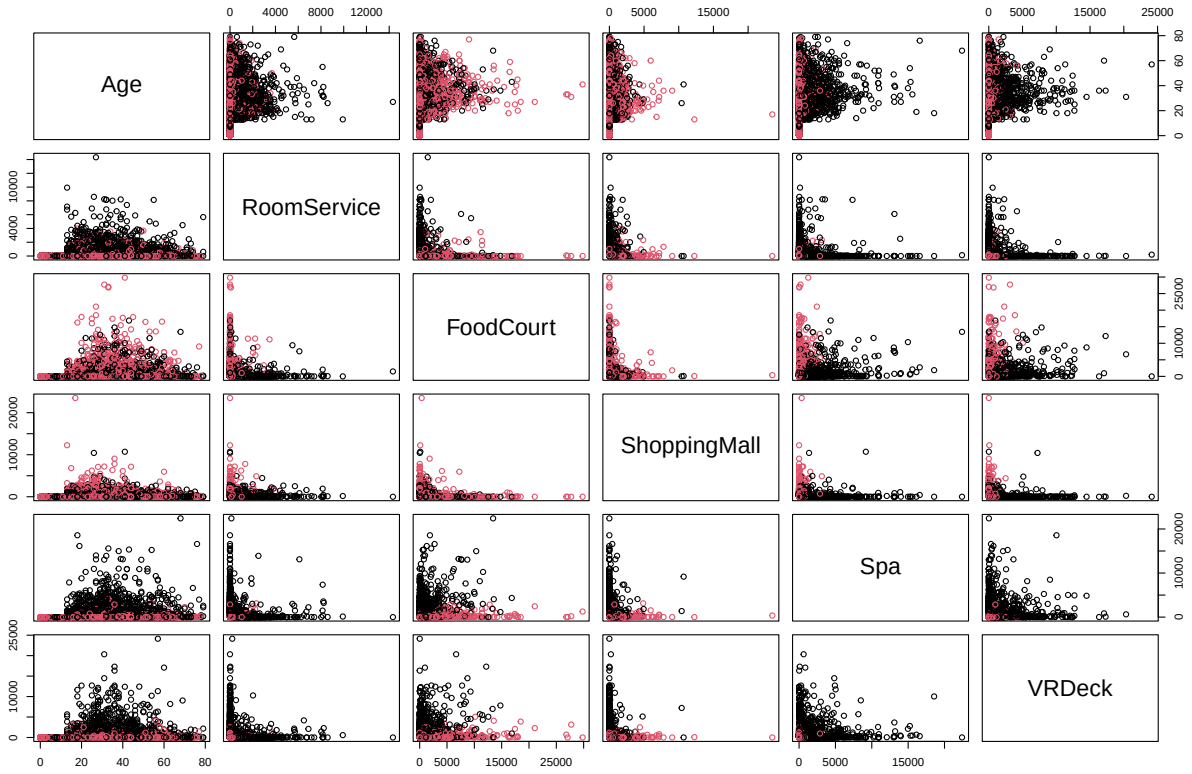

Figure 5: Relationships Between Variables: Red Points are Transported

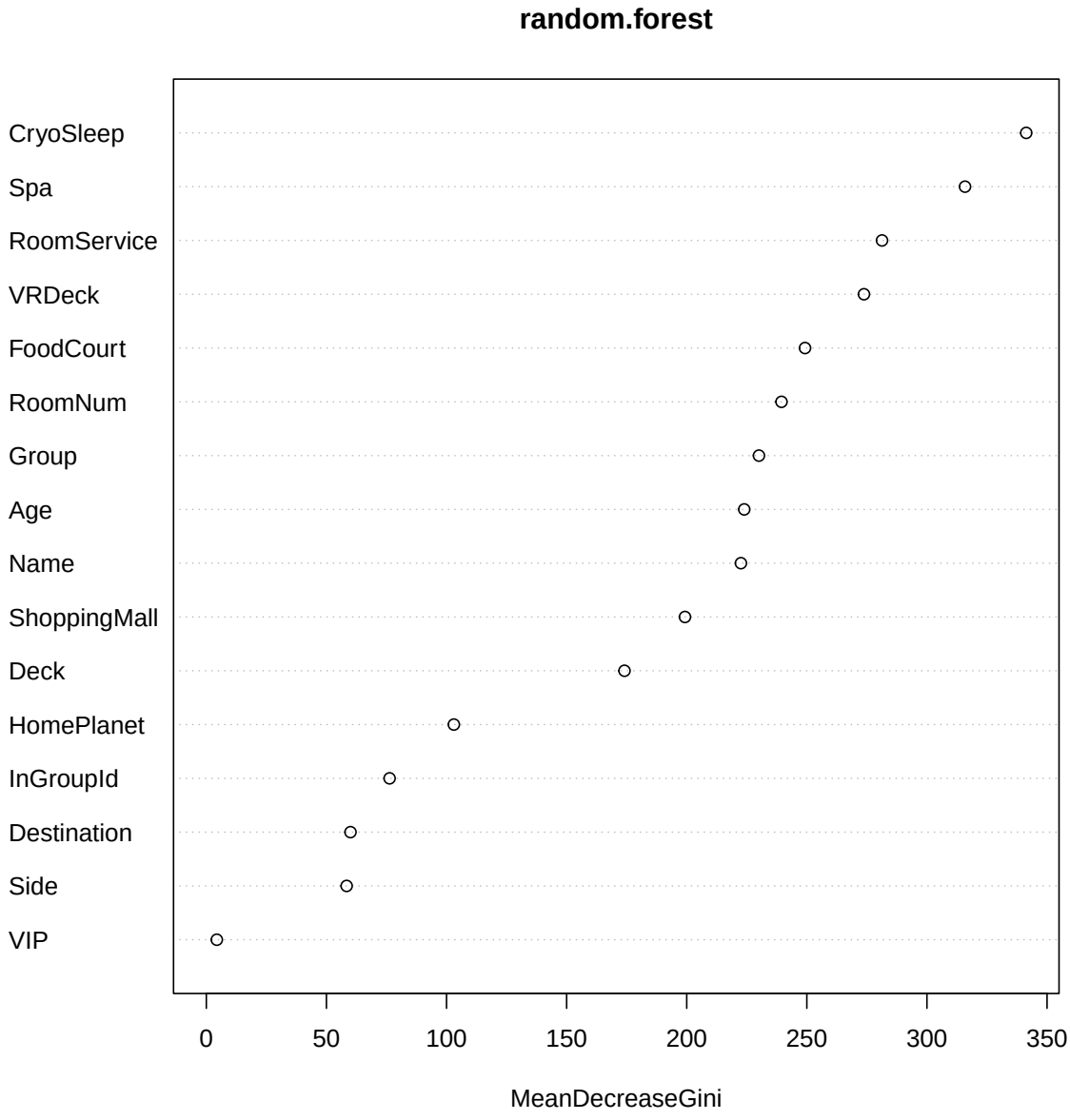

Figure 6: Variable Importance