Spaceship Titanic Project

Authors: Monica Amezquita; Alejandro Hernandez; Hugo Marquez
Date Submitted: August 10th, 2022

I. Problem Description

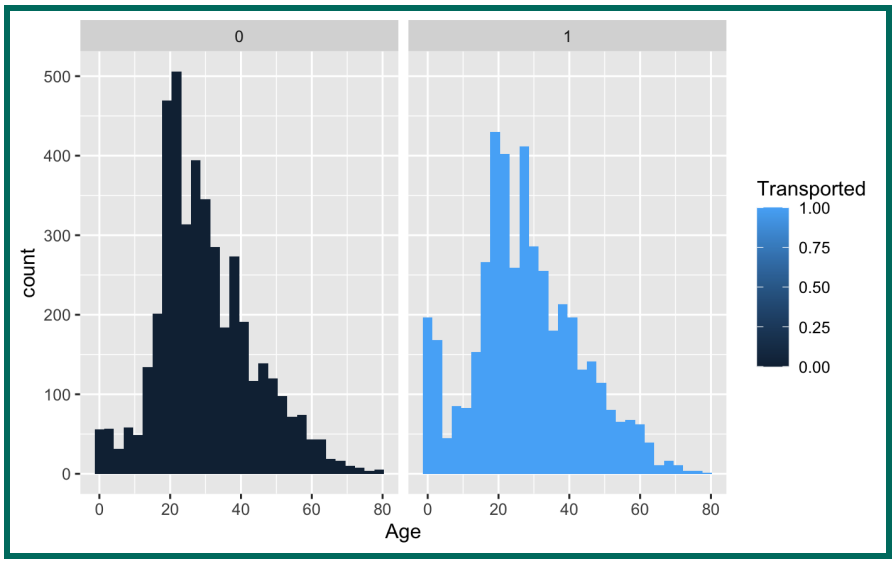
In the year 2912, the *Spaceship Titanic* was an interstellar passenger liner with over 13,000 individuals on board. The passenger liner was en route to transport emigrants from Earth, Mars, and Europa to three newly discovered exoplanets: TRAPPIST-1e, PSO J318.5-22, and 55 Cancri E. While the Spaceship Titanic was on its way to the first destination, the interstellar liner collided with a spacetime anomaly which resulted in almost half of the passengers being transported to an alternate dimension. This project seeks to predict which passengers were transported using the dataset provided.

II. Exploratory Analysis

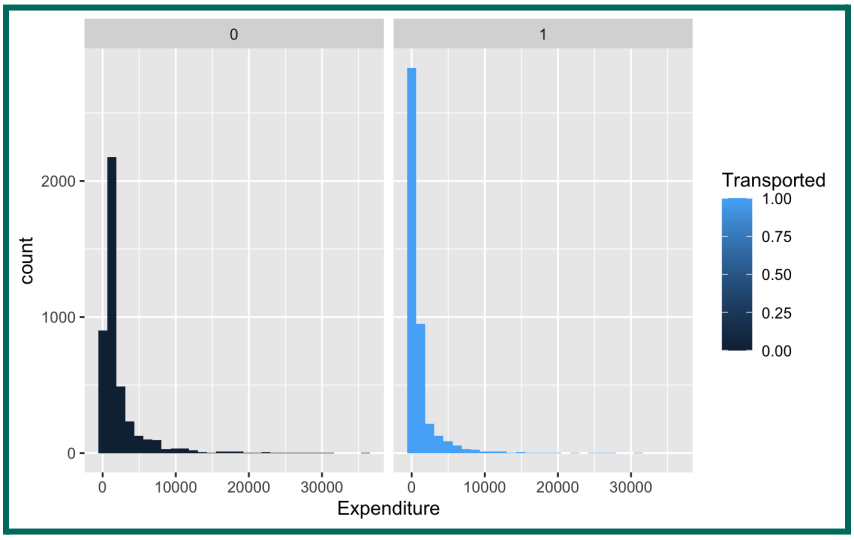
The Spaceship Titanic dataset consists of thirteen predictors and one response variable. The first are a collection of passenger attributes: *PassengerId**,** HomePlanet, Cryosleep, Cabin (Deck/RoomNumber/Side), Destination, Age, VIP, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck,* and *Name*. *Transported* is the response variable which consists of boolean data, representing whether or not an individual was transported. The provided dataset (for training and model validation) is well balanced consisting of 49.64% of individuals who were not transported and the remaining 50.36% individuals were transported, for a total of 8,693 observations.

Basic data visualizations uncover relationships between *CryoSleep* and *Transported*, expenditure variables and *Transported*, as well as *Deck* and *Transported*. It is important to note that the relationships between certain features are likely to be correlated with other features. For example, *CryoSleep* and total expenditure are highly correlated as individuals who opted for cryosleep did not make any purchases on the Spaceship Titanic, and therefore, had zero expenditures. Furthermore, *Age* also has an impact on total expenditure as children did not have expenditures (or spent very minimal amounts), and older individuals had higher expenditures. In addition, expenditures within groups also had similar spending habits.

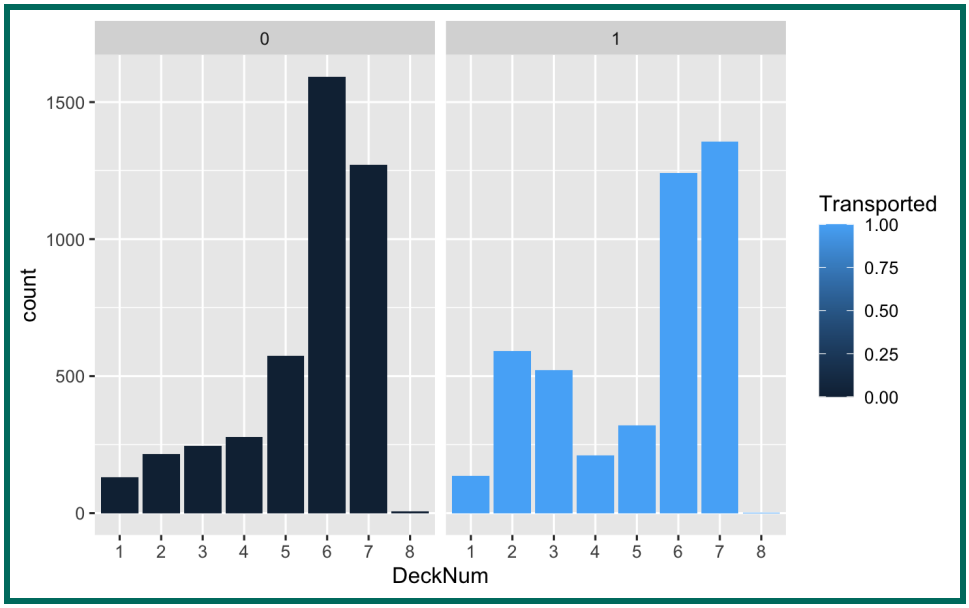
In the following plots, we seek to analyze the relationships between transported and variables such as age, expenditure, and deck. The following histogram of *Age* colored by *Transported* shows that babies had a higher chance of being transported, while young adults around 19 years old show a decreased chance of being transported.



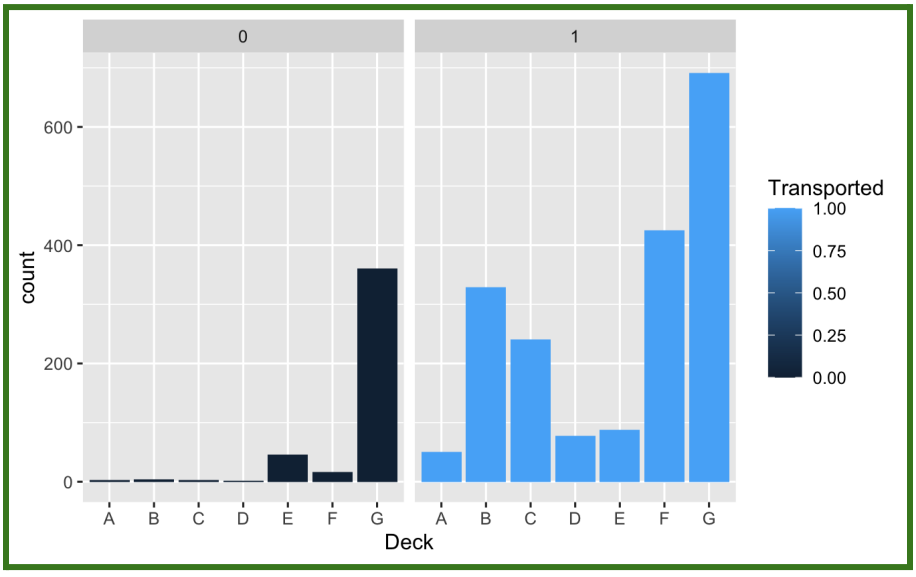
The following plot shows that individuals with zero expenditure had a higher chance of being transported while individuals with some expenditure show lower rates of transportation.



Finally, we explored the relationship between a passenger's deck and whether or not they were transported. We can see that there are higher rates of transportation in decks 2,3, and 7 and there were lower rates of transportation in decks 5 and six.



While these plots show that there are relationships between the aforementioned predictor variables, we cannot say that these variables are responsible for the rates of transportation. There can be hidden relationships between these variables and other predictor variables. One predictor variable that comes to mind is *CryoSleep*. As previously mentioned, cryosleep status affected the expenditure variables, but it can also be related to deck location. It is possible that the passenger liner places all cryosleepers in certain areas of the ship. The following plot is a bar chart showing the frequencies of cryosleepers per deck colored by whether or not they were transported. We can see that almost all cryosleepers were transported, with only decks E and G showing greater amounts of cryosleepers that were not transported.



In decks A, B, C, and D, there were only between two and four cryo-sleepers that did not get transported, while in decks E, F, and G, there were 46, 17, and 361. We can see the values in the following table:

	Deck	count
1	A	3
2	B	4
3	C	3
4	D	2
5	E	46
6	F	17
7	G	361

The following table shows the frequencies of individuals that did get transported in their respective decks:

	Deck	count	Transported
1	A	50	1
2	B	329	1
3	C	241	1
4	D	78	1
5	E	88	1
6	F	425	1
7	G	691	1

III. Data Pre-Processing

Pre-processing is a necessary step before model building for most datasets- and this is no exception. Below is a description of our methods for estimating missing values and engineering new predictor variables.

A. Data Imputation

Imputation is the practice of estimating a missing value of an observation/record by inference or prediction. There are non-parametric (KNN, mean/mode, etc.) and parametric (Linear Regression, LDA etc.) methods. In this project, we utilized three main methods for imputation missing data, two of which are uniquely motivated by the dataset.

Two important notes: First, missing data was present in our training, validation, and testing sets. One may assume it acceptable to estimate the values of each independently (e.g. use validation data to estimate missing values in the validation set), although that would be incorrect. To prevent data leakage, information from the validation and testing sets must be "locked away" up until the point they are used for evaluation. Therefore, any model used to estimate missing data in the training/validation/testing sets must be fitted from the training data alone. Second, of any predictor with incomplete data (for any of the data sets), less than 2% was missing. Because our imputed values were such a small percentage of any data set, we don't expect the following imputation methods to have a significant impact on future performance.

1. Group-based Imputation

The intuition behind this method is that values of certain variables are similar among members of a group. For example, we would expect members of a group to have a similar social status, similar spending habits, etc. Then, we can utilize mean/mode values for a variable within a group, and use it as an estimate of missing-data-members. In selecting predictors eligible for this method, we conducted simple analysis to identify patterns among groups; for more information, see Supplementary.

2. Planet-based Imputation

Consider a solo passenger who is missing information about their home or destination planet. We cannot rely on other group members to estimate these values. Fortunately, we can analyze these variables over the entire training set and identify the most common destination from a given starting point, and vice versa. This describes our imputation method to estimate the missing values of solo riders for the variables *HomePlanet* and *Destination*.

3. Multivariate Imputation

For all remaining variables with missing values, we fit a multivariate parametric model according to its type: lasso regression for numeric, logistic regression with lasso penalty for binary, and LDA for nominal. These models were trained using the Multivariate Imputation by Chained Equations (MICE) library. For numeric variables that encode attributes like age and currency spent at various facilities, the lasso regression model sometimes returned negative values. We interpreted this output as the model predicting these values to be so low as to be negative. To handle this issue, we converted all negative imputations to zero.

Another method we explored was the general mean/mode method, but due to its infamy and crudeness, ultimately we elected to avoid it. Read more in Supplementary.

B. Feature Engineering

Feature engineering is the process of using existing data to create new variables that are not currently in the training dataset. Creating new features can be extremely useful towards improving the accuracy of the models on both training and testing datasets. Various features within the Spaceship Titanic dataset contain information that is more useful when transformed. The *PassengerId* feature is of the form “gggg_pp”, where “gggg” is an individual’s group number and “pp” is an individual’s number within a group. By splitting the *PassengerId* feature into a *GroupId* and *PersonalId*, we can analyze whether individuals in groups are more or less likely to be transported. We further split Cabin into more informative features - *Deck* and *Side*. Deck held nominal categorical data which we converted to an ordinal categorical feature, ranking the different levels of the ship. *Side* was converted to a binary categorical feature where a zero represents “port side” passengers and a one represents “starboard” passengers. We further convert the *HomePlanet* variable to a binary categorical feature where Europa is represented by a zero and both Earth and Mars are represented by a one. Earth and Mars are represented by the same level due to the closer distance between these planets, while Europa is the home planet out of the possible options. Lastly, we created an *Expenditure* feature that consists of summing all expenditure

features like *Spa, ShoppingMall*, and *VRDeck*. Due to the relationship between features like cryosleep and expenditure, some models had issues with having too many ties such as in the case of KNN. To reduce the number of ties, we added a small normal random amount to all expenditure features which fixed this issue.

IV. Model Building and Evaluation

Based upon the amount of observations in this data set (8,693) we split the data into two sets. The split will consist of the training set and validation set with a split percentage of 75% and 25% respectively. In this partition our response variable will consistently be *Transported* for those who were transported to an alternate dimension.

In order to predict the response we must use and implement appropriate classification models. We will be classifying whether someone was *Transported* to an alternate dimension based upon the personal records recovered by the damaged systems for those on board. Furthermore, levels were added for those who were transported, defined as “No” or “Yes”.

Predictors used to predict Transported (response variable):

- **People_in_group** (if they were traveling as a group)
- **GroupID**
- **HomePlanet** (Earth, Mars, Europa)
- **CryoSleep** (if a person was sleeping throughout the trip)
- **Side** (spaceship side)
- **Destination** (one of the three exoplanets they were traveling to)
- **Age**
- **Expenditure** (the amount of money passengers spent while on board)
- **DeckNum (**the level/ deck passengers were in)

A. Classification Models Used

Logistic Regression:

This particular model allowed the use of the “glm” method in practice when fitting our predictors. Furthermore, this model is beneficial since we have a binary response variable. Taking precaution when using this model was taken into great consideration but with the use of feature engineering the amount relevant predictors was low and overfitting was not a concern.

K-Nearest Neighbors:

This model being used is non-parametric since it doesn’t have predefined predictors unlike linear or logistical methods. This method used a cross validation of 20 since our data was large. Furthermore, this method estimates the likelihood of those who were transported by evaluating the nearby data points.

Lasso classification:

This model used the "glmnet" method approach where we kept the same predictors used in the previous models to predict the binary response variable *Transported*. This model allows for interpretability since we are not using a large number of variables and will also shrink coefficients to zero. Lasso here has an alpha = 1 to produce as many coefficients as possible when they are getting zeroed out. Also, taking into consideration the usefulness of this model because it does not penalize for adding more predictors or interactions into its fit.

Ridge Classification:

Ridge Classification is also being generalized using the "glmnet" method similar to Lasso, we are keeping the same predictors but since we are using Ridge it needs alpha = 0. Unlike Lasso, Ridge will have non-zero coefficients and require a tuning parameter to find the minimum since it was not hitting it when lambda =1. Similarly, Ridge classification does not penalize or result in overfitting the model when additional predictors are added therefore, we used eight predictors since the size of our data also allowed us to take advantage of this condition.

Linear SVM:

Linear SVM is a beneficial method where it allows us to make a decision boundary for those who were transported to an alternate dimension. This algorithm allows for the data to be split into classes and in our case we have n dimensions since we have multiple predictors. This will create visualization challenges since we are working with a high dimension which will result in a hyperplane once we get past the second dimension. It is important to notice the cutoff because it will be needed as probabilistic outcomes when producing the boundary plots.

Polynomial SVM:

Similar to Linear SVM it is beneficial but more flexible since we are working with a higher degree polynomial. This model will produce up to a third degree polynomial where the second and third degree are more flexible and better split our data.

RBF SVM:

RBF SVM works similar to polynomial SVM where it works to find non-linear classifiers. Like the other model we must be careful of overfitting and over estimating therefore we must keep in mind the boundary and how wide the support vectors are.

Fully Connected Neural Network:

Using 15 nodes to construct a neural network using the ReLU activation function this will help us predict those who were transported. The benefits of this model consist of its connectedness and how the nodes are connected to all the other nodes. This occurs for as many layers as you have in the model while the ending result consists of the data provided by the previous layers so that the ending layer (output layer) gives the final output.

B. Model Methodology

Model selection was approached with caution, interpreting the models and what they can provide for us as we fit each one . Understanding the model capabilities and avoiding overfitting models such as KNN and Logistic was well compensated for in our selection of predictors. Through the use of feature engineering and imputation we established a strong understatement of our predictors which then helped us interpret the influence they had on the response variable (*Transported*). Focusing on the predictors influence we then decided to only use those that we thought were of great use. Using more predictors could have caused issues such as multicollinearity because one predictor could be predicted from another. The collinearity issue was visible in predictors such as *Expenditures, FoodCourt, and RoomService* where they are related to the amount of money spent on the cruise therefore having them all together in a model would have resulted in a linear relationship of some degree when evaluating the models. Therefore, focusing on having a small number of predictors did not cause issues in our model fit although models such as LASSO and Ridge could have benefited from the kitchen sink approach where we use all predictors since there is no penalty. The model selection was precise since we had a classification problem therefore, we couldn't use linear regression models since our response variable was not numerical and instead a class. We decided to keep the same predictors for all models (except Neural Network) and compare the model's results in the following section through the use of ROC curves and accuracy rates.

C. Model Results

The models tested in this project produced very similar results. On average, all models produced about 72% accuracy, with lasso and ridge performing the best at 73.56% and 73.76% accuracy respectively. Because the lasso model zeros out coefficients that were not significant in predicting the teleported individuals, we obtained a better accuracy score compared to models like KNN which used all features included in the model, regardless of whether or not the features were significant. The lasso model produced the following coefficients:

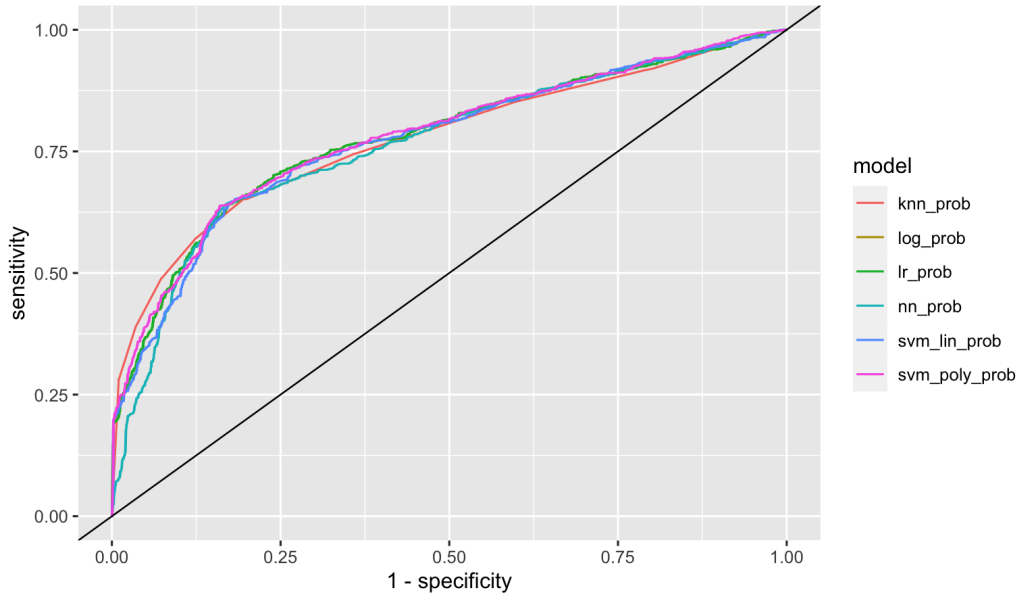
```
10 x 1 sparse Matrix of class "dgCMatrix"
                                      s1
(Intercept)              -0.07566701
HomePlanet               -0.34892593
CryoSleepTRUE             0.99299837
DeckNum                  -0.01132919
Side                      0.06005447
DestinationPSO J318.5-22  .
DestinationTRAPPIST-1e    .
Age                       .
Expenditure              -0.03925901
people_in_group           .
```

We can see that *HomePlanet, CryoSleepTrue, DeckNum, Side,* and *Expenditure* were the most significant coefficients for the model, with *CryoSleepTrue* having the greatest weight. While this model gave us a high accuracy, we could have improved the model by increasing the number of features used as well as introducing interaction terms and allowing lasso to zero out any unnecessary coefficients.

Polynomial support vector machine and RBF support vector machine also performed well on both the training set and validation set, producing an accuracy of 73.29% and 73.20% accuracy respectively.

These models were more challenging to work with due to high run times. It was difficult to test the models with different features as each run was extremely time consuming. In the future, we could try running a lasso model to determine which coefficients could be helpful, and using these features on a SVM model.

The remaining KNN, logistic regression, and linear SVM models all performed similarly with an accuracy of about 72%. We could improve these models by changing the features and introducing more robust features during the feature engineering process. The following plot shows the different ROC curves for all models on the validation set:



As per the plot, we can see that all models performed about the same on the validation set. All models had a greater true negative rate (determining individuals that were not transported) compared to the true positive rate (determining individuals that were transported). Furthermore, all models produced higher amounts of false negatives compared to the false positive amounts. The following table is the confusion matrix for the neural network, which is highly representative the confusion matrices for all models:

	Truth	
Prediction	No	Yes
No	892	394
Yes	186	700

A model that predicts false negatives at greater rates than the prediction of false positives can be a dangerous model as we can incorrectly tell an individual that they will not be teleported even though they will be transported. A model that has greater false positive rates is less hazardous to passengers as these individuals will not be transported anyways, so whether or not they choose to board the ship will not make any difference to their overall safety. For these reasons, we must continue working on a model that has greater true positive rates and lower false negative rates. This can be accomplished through further

extensive data exploratory practices, further feature engineering and selection, model optimization, and possibly enhanced data imputation methods.

V. Final Model and Kaggle Score

As discussed in the model results section, lasso and ridge performed the best on the metric of accuracy, with ridge producing a higher accuracy rate. The final ridge model used on the test dataset produced the following coefficients:

```
10 x 1 sparse Matrix of class "dgCMatrix"
                                      s1
(Intercept)                0.208721758
HomePlanet                -0.209318260
CryoSleepTRUE              0.506066834
DeckNum                   -0.036960376
Side                       0.130975935
DestinationPSO J318.5-22 -0.043110169
DestinationTRAPPIST-1e   -0.090523370
Age                       -0.002719505
Expenditure               -0.021834912
people_in_group            0.017491198
```

While the ridge is less interpretable than the lasso model due to the use of all coefficients, it produced the highest accuracy of all models. Therefore, we predicted on Kaggle's test dataset using the trained ridge model. Our predictions produced an accuracy score of 73.76% on Kaggle, which is indicative of the performance produced on the validation dataset.

VI. Conclusion and Future Work

The Spaceship Titanic dataset provides an opportunity to develop statistical models on a binary classification dataset. The dataset provides a smaller number of features which further lends the opportunity to develop new features during the feature engineering process. After developing numerous models including logistic regression, lasso and ridge, KNN, SVM (linear, polynomial, and RBF), and a one layer neural network, we found that lasso and ridge performed the best on the features included in the model. Lasso performed well due to the elimination of any insignificant coefficients, while ridge performed well due to the high correlation between features such as expenditure and cryosleep. After predicting on the test dataset using the ridge model, we obtained a 73.76% accuracy score on Kaggle.

Like all projects, there is certainly room for improvement in our process. Our model's accuracy can be improved through the development of new features during the feature engineering process and further data exploratory analysis to further understand relationships. Additionally, an oversight of our methodology was not including interaction terms in our models. Models built from a mostly categorical dataset are sure to be improved with the addition of interaction terms. With such additions to our project, models can be optimal in regards to our response variable. In the future, taking advantage of lasso's properties and its flexibility to add a large amount of predictors with no penalty can be extremely beneficial in our efforts to find those passengers who were transported to an alternate dimension in this unforeseeable event.

VII. Supplementary

A. Selecting Variables Eligible for Group-Based Evaluation

Not all variables were assumed to have similar values among members of a passenger group. This section details our testing of the expenditure-related variables as eligible for the group-based imputation method. The process of this test was to first compute the average value of a certain variable, within each group with more than one member. Then, for each of these groups, we can calculate that predictor's variance and standard deviation. Finally, we average those variance and standard deviations across all groups. Below is the (rounded) average group variance and standard deviation, across all groups, for the given variables.

	Mean Group Variance	Mean Standard Deviation
Room Service	482,146	**276**
Food Court	3,470,367	742
Shopping Mall	457,243	**230**
Spa	1,528,962	492
VR Deck	1,722,467	495

This table gives us a general idea of how groups deviate from their variable means. Some variables have a high average group deviation from the mean, such as *FoodCourt*, making a method like group-based imputation less successful for this variable; recall that the group-based method imputes missing data with group averages.

Then, we repeated this test for only groups with missing data. After all, the estimated averages we use to impute missing data are from such groups, so it may be helpful to see if any variable has a lower mean standard deviation among missing-data groups.

	Mean Group Variance	Mean Standard Deviation
Room Service	329,881	**235**
Food Court	2,864,001	795
Shopping Mall	61,437	**115**
Spa	1,612,618	602
VR Deck	1,724,427	613

As we can see, the average group standard deviation is quite low for the variables *RoomService* and *ShoppingMall*. If the group values of *ShoppingMall* differ by only 115 units of currency, using the average as an estimate for unknown-group-member data may be a fine estimate. These metrics directed us to impute those predictors with the group-based method.

B. Mean/Mode Imputation

Filling missing values of a variable with the variable's overall mean (if numeric) or mode (if categorical) is a simple strategy that seems- at a glance- reasonable. However, there are some serious drawbacks to this method. Because these imputed values are totally independent from other predictors, they have the potential to erase correlation between variables. This is certainly an issue when a significant chunk of the data is missing. In our case, despite such few incomplete records, we decided there are more motivated and refined methods of imputation.