Homework 3

PSTAT 134/234

Table of contents

Homework 3

For this homework assignment, we'll be working with a dataset called Spaceship Titanic. It is a simulated dataset used for a popular Kaggle competition, intended to be similar to the very famous Titanic dataset. The premise of the Spaceship Titanic data is that it is currently the year 2912. You have received a transmission from four lightyears away, sent by the Spaceship Titanic, which was launched a month ago.

The Titanic set out with about 13,000 passengers who were emigrating from our solar system to three newly habitable exoplanets. However, it collided with a spacetime anomaly hidden

Figure 1: Star Trek ships.

in a dust cloud, and as a result, although the ship remained intact, half of the passengers on board were transported to an alternate dimension. Your challenge is to predict which passengers were transported, using records recovered from the spaceship's damaged computer system.

The dataset is provided in `/data`, along with a codebook describing each variable. You should read the dataset into your preferred coding language (R or Python) and familiarize yourself with each variable.

We will use this dataset for the purposes of practicing our data visualization and feature engineering skills.

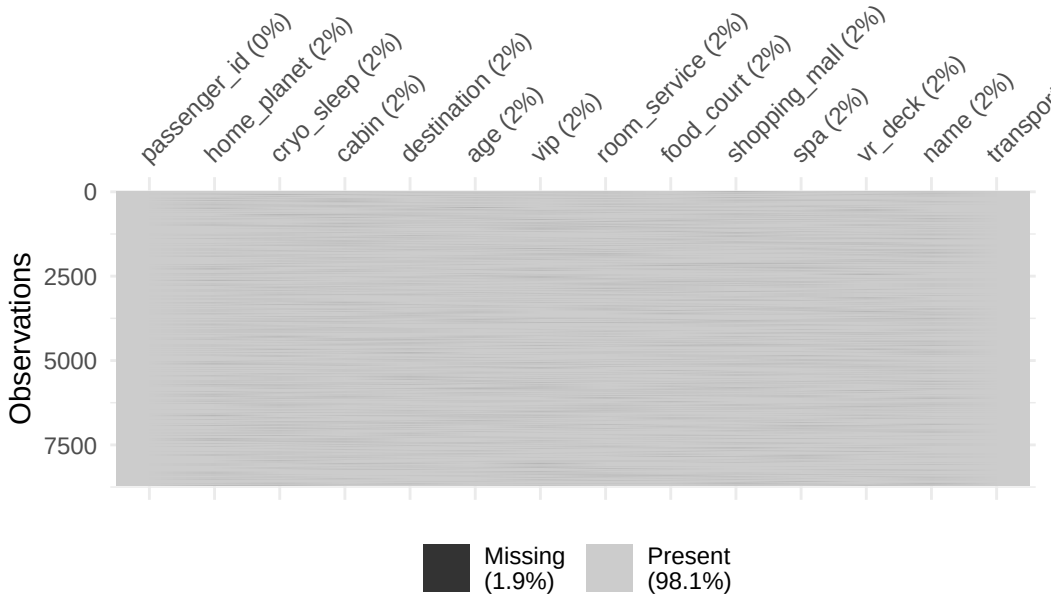
Exercise 1

Which variables have missing values? What percentage of these variables is missing? What percentage of the overall dataset is missing?

```
library(tidyverse)
library(naniar)
library(janitor)
library(imputeTS)
library(ggplot2)
```

```r
library(corrplot)
library(recipes)

spaceship_data <- read.csv("./data/spaceship_titanic.csv") %>%
  mutate_all(~na_if(as.character(.), "")) %>% clean_names()
vis_miss(spaceship_data)
```



All variables except passenger_id and transported have missing values, and all of these variables with missing values have about 2% of missing data. The overall dataset has 1.9% missing values.

Exercise 2

Use mode imputation to fill in any missing values of `home_planet`, `cryo_sleep`, `destination`, and `vip`. Drop any observations with a missing value of `cabin` (there are too many possible values).

```r
# mode function
impute_mode <- function(x) {
  mode_val <- names(sort(table(x), decreasing = T))[1] # find mode
  x[is.na(x)] <- mode_val # replace NA
  return(x)
}
```

```r
# mode imputation & filter missing value of cabin
spaceship_data <- spaceship_data %>%
  mutate(home_planet = impute_mode(home_planet),
         cryo_sleep = impute_mode(cryo_sleep),
         destination = impute_mode(destination),
         vip = impute_mode(vip)) %>%
  filter(!is.na(cabin))
```

Use median imputation to fill in any missing values of `age`. Rather than imputing with the overall mean of `age`, impute with the median age of the corresponding `vip` group. (For example, if someone who is a VIP is missing their age, replace their missing age value with the median age of all passengers who are **also** VIPs).

```r
median_vip <- median(as.numeric(spaceship_data$age[spaceship_data$vip == "True"]), na.rm =
median_non_vip <- median(as.numeric(spaceship_data$age[spaceship_data$vip == "False"]), na

spaceship_data <- spaceship_data %>%
  mutate(age = ifelse((vip == "True" & is.na(age)), median_vip, age)) %>%
  mutate(age = ifelse((vip == "False" & is.na(age)), median_non_vip, age))
```

For passengers missing any of the expenditure variables (`room_service`, `food_court`, `shopping_mall`, `spa`, or `vr_deck`), handle them in this way:

- If all their observed expenditure values are 0, **or** if they are in cryo-sleep, replace their missing value(s) with 0.
- For the remaining missing expenditure values, use mean imputation.

```r
room_service_mean <- mean(as.numeric(spaceship_data$room_service), na.rm = T)
food_court_mean <- mean(as.numeric(spaceship_data$food_court), na.rm = T)
shopping_mall_mean <- mean(as.numeric(spaceship_data$shopping_mall), na.rm = T)
spa_mean <- mean(as.numeric(spaceship_data$spa), na.rm = T)
vr_deck_mean <- mean(as.numeric(spaceship_data$vr_deck), na.rm = T)

spaceship_data <- spaceship_data %>%
  mutate(across(c(room_service, food_court, shopping_mall, spa, vr_deck), as.numeric)) %>%
  mutate(across(c(room_service, food_court, shopping_mall, spa, vr_deck),
                ~ ifelse(is.na(.) & (cryo_sleep == "True" |
                                     rowSums(cbind(room_service, food_court, shopping_mall
  mutate(
    room_service = ifelse(is.na(room_service), room_service_mean, room_service),
    food_court = ifelse(is.na(food_court), food_court_mean, food_court),
```

```
    shopping_mall = ifelse(is.na(shopping_mall), shopping_mall_mean, shopping_mall),
    spa = ifelse(is.na(spa), spa_mean, spa),
    vr_deck = ifelse(is.na(vr_deck), vr_deck_mean, vr_deck)
  )
```

Exercise 3

What are the proportions of both levels of the outcome variable, `transported`, in the training set?

```
spaceship_data %>%
  count(transported) %>%
  mutate(proportion = n / sum(n))
```

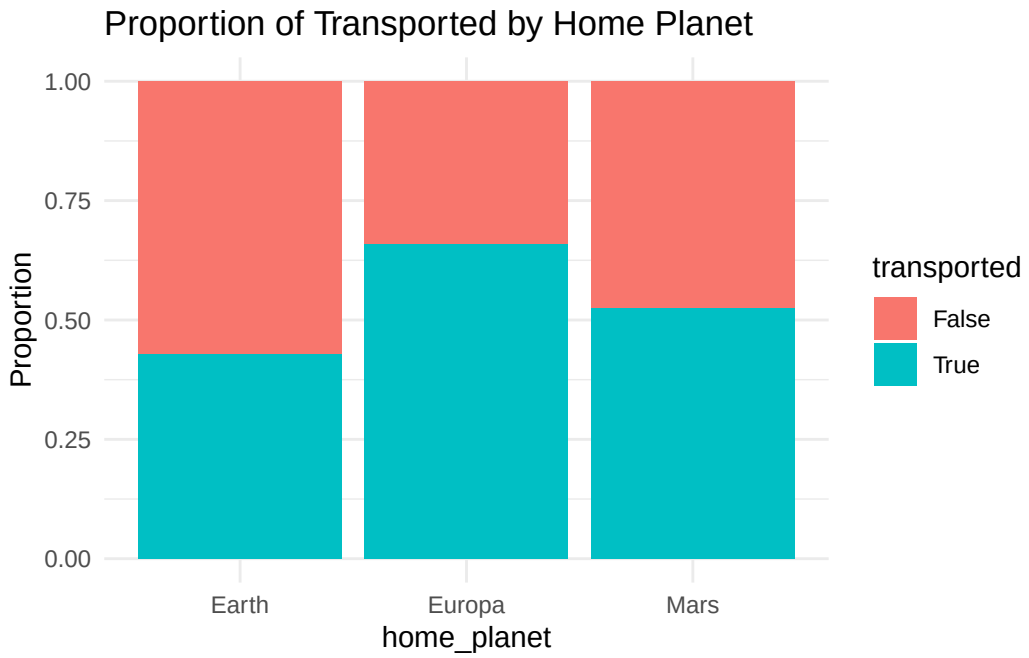
```
  transported    n proportion
1       False 4216  0.4963504
2        True 4278  0.5036496
```

Exercise 4

Make proportion stacked bar charts of each of the following. Describe what patterns, if any, you observe.

1. `home_planet` and `transported`

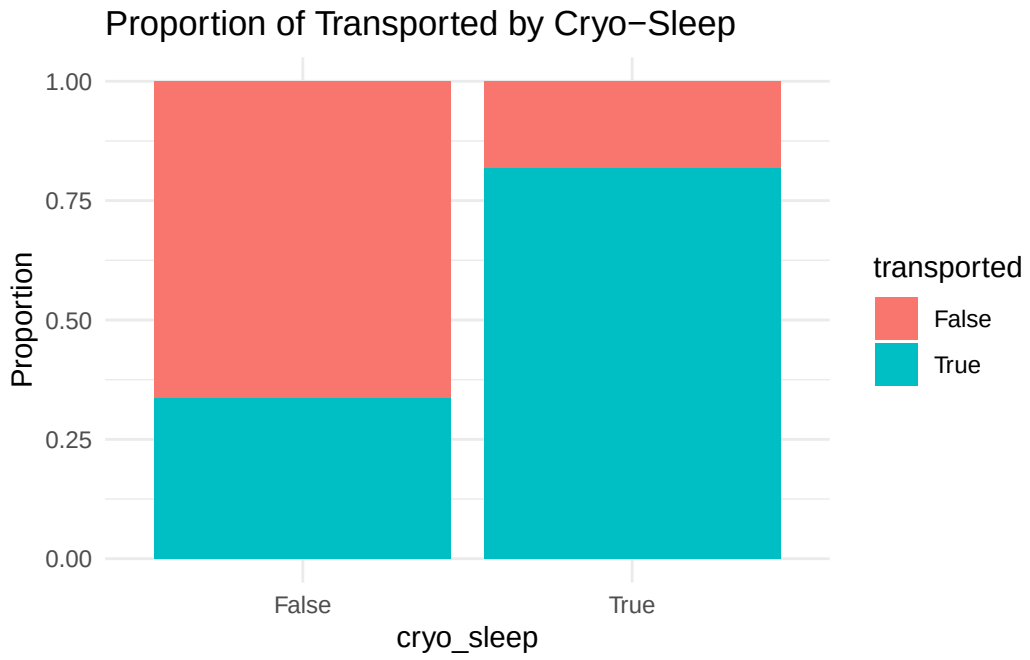
```
ggplot(spaceship_data, aes(x = home_planet, fill = transported)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion", title = "Proportion of Transported by Home Planet") +
  theme_minimal()
```



More passenger was transported to another dimension if their home planet is Europa.

2. `cryo_sleep` and `transported`

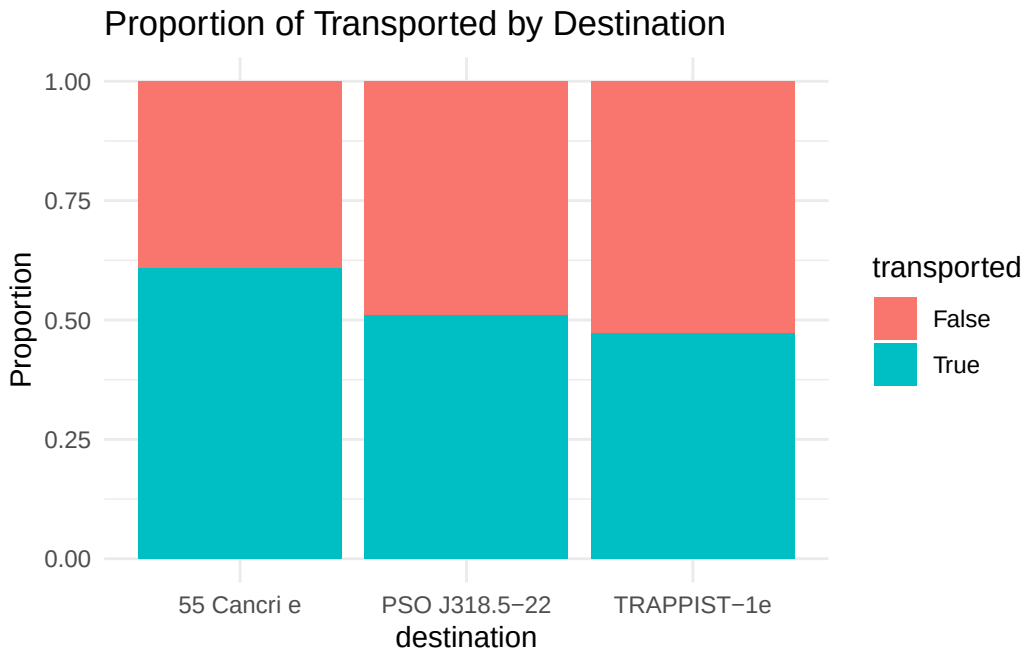
```
ggplot(spaceship_data, aes(x = cryo_sleep, fill = transported)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion", title = "Proportion of Transported by Cryo-Sleep") +
  theme_minimal()
```



More passenger was transported to another dimension if Cryo_sleep stage is True.

3. `destination` and `transported`

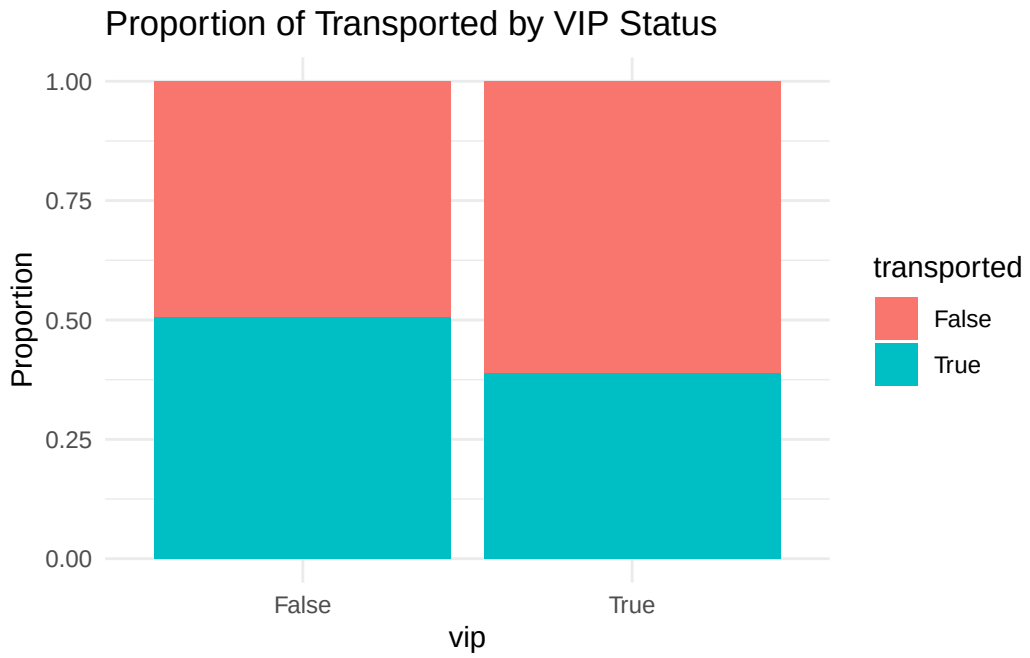
```
ggplot(spaceship_data, aes(x = destination, fill = transported)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion", title = "Proportion of Transported by Destination") +
  theme_minimal()
```



More passenger was transported to another dimension if their destination is 55 Cancri e.

4. `vip` and `transported`

```
ggplot(spaceship_data, aes(x = vip, fill = transported)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion", title = "Proportion of Transported by VIP Status") +
  theme_minimal()
```



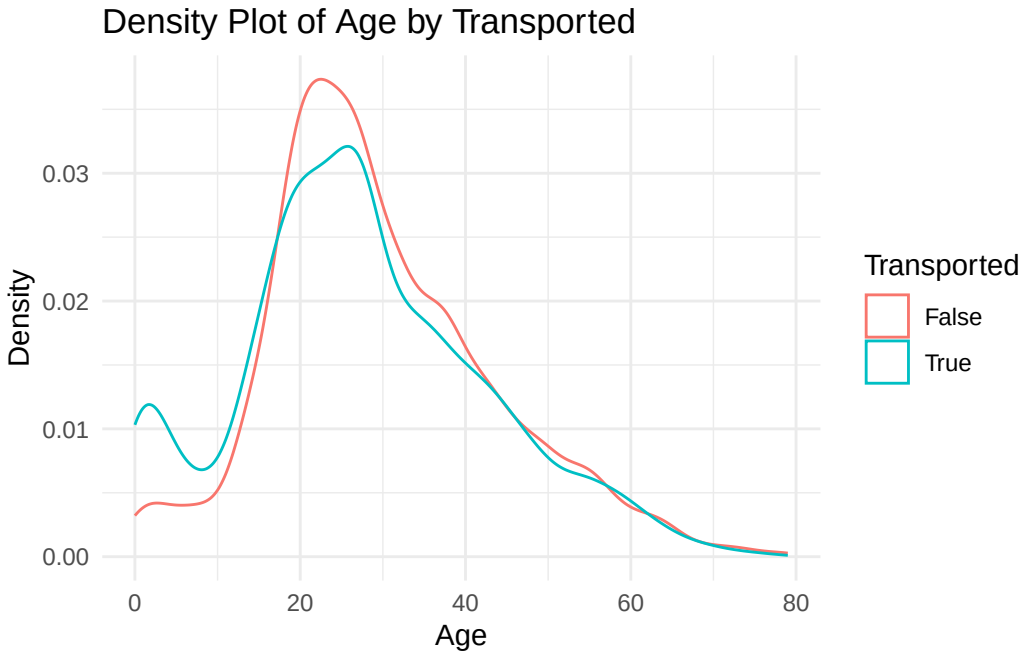
More passenger was transported to another dimension if they are not VIP.

Exercise 5

Using box plots, density curves, histograms, dot plots, or violin plots, compare the distributions of the following and describe what patterns you observe, if any.

1. `age` across levels of `transported`

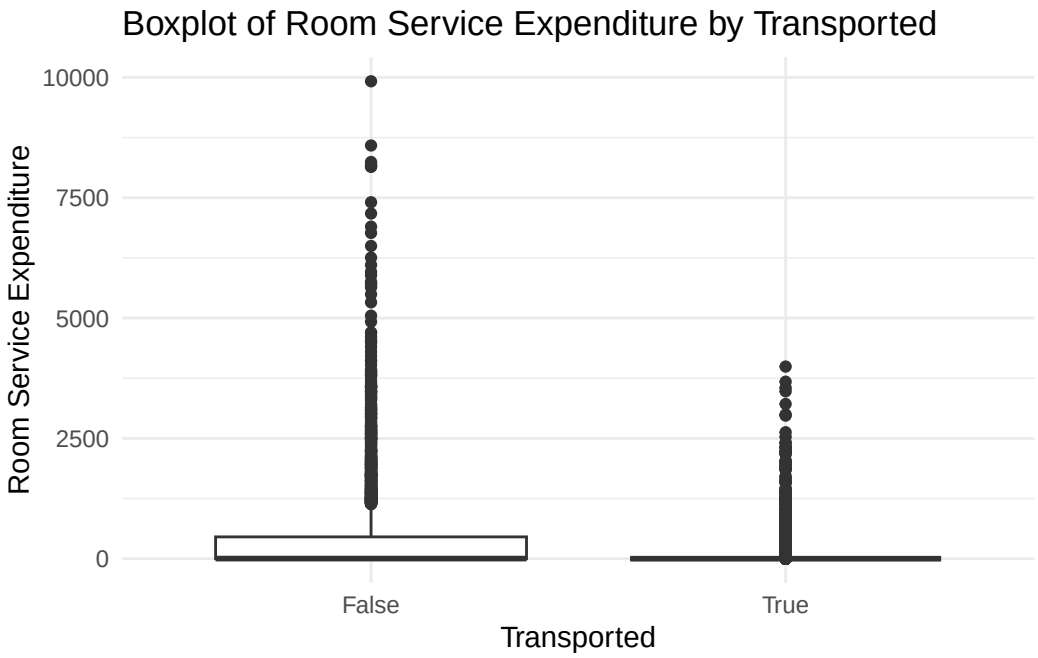
```
ggplot(spaceship_data, aes(x = as.numeric(age), color = factor(transported))) +
  geom_density() +
  labs(title = "Density Plot of Age by Transported",
       x = "Age",
       y = "Density",
       color = "Transported") +
  theme_minimal()
```



Younger passengers who have a age below 20 are more likely to be transported. And passengers around the age of 20-30 are less likely to get transported.

2. `room_service` across levels of `transported`

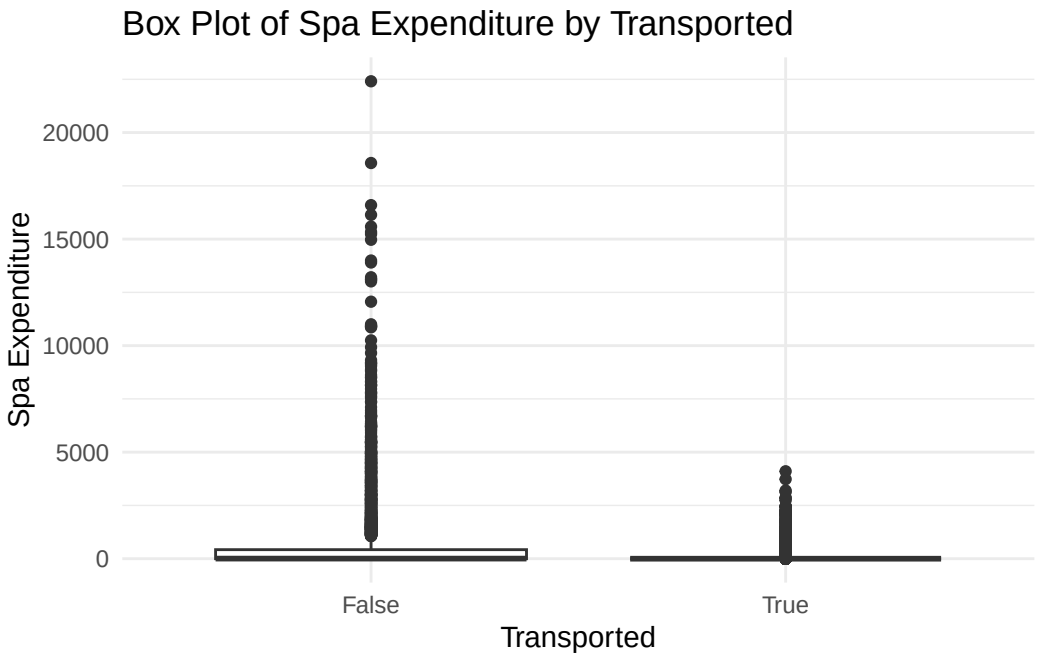
```
ggplot(spaceship_data, aes(x = transported, y = room_service)) +
  geom_boxplot() +
  labs(title = "Boxplot of Room Service Expenditure by Transported",
       x = "Transported",
       y = "Room Service Expenditure") +
  theme_minimal()
```



Passengers are more likely to not be transported if they spend a large amount of money on room service.

3. `spa` across levels of `transported`

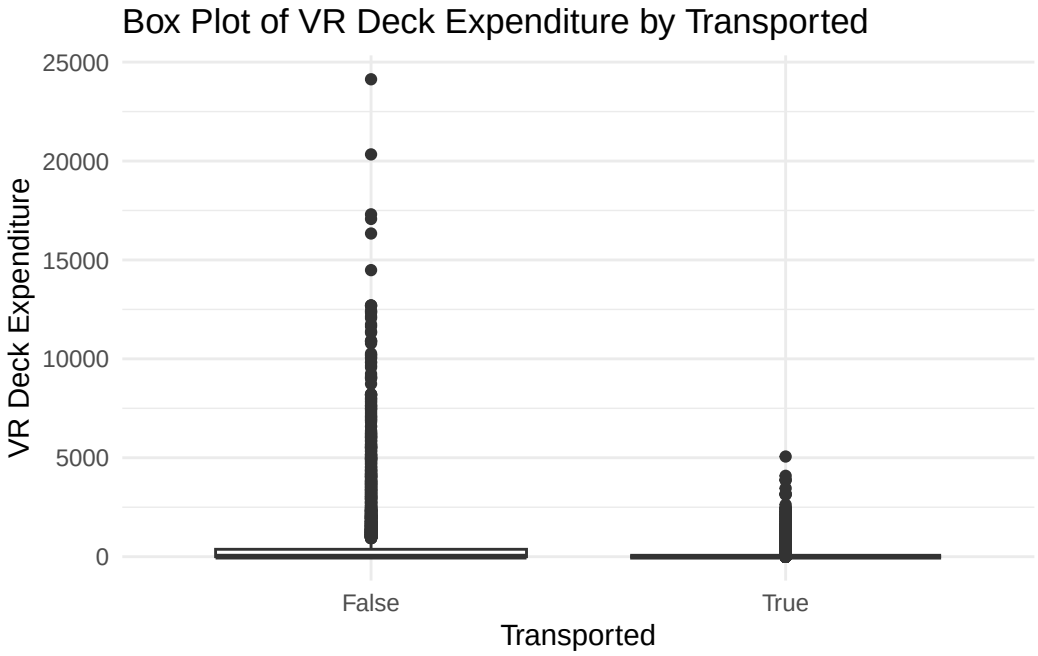
```
ggplot(spaceship_data, aes(x = transported, y = spa)) +
  geom_boxplot() +
  labs(title = "Box Plot of Spa Expenditure by Transported",
       x = "Transported",
       y = "Spa Expenditure") +
  theme_minimal()
```



Passengers are more likely to not be transported if they a large amount of money on Spa.

4. `vr_deck` across levels of `transported`

```
ggplot(spaceship_data, aes(x = transported, y = vr_deck)) +
  geom_boxplot() +
  labs(title = "Box Plot of VR Deck Expenditure by Transported",
       x = "Transported",
       y = "VR Deck Expenditure") +
  theme_minimal()
```



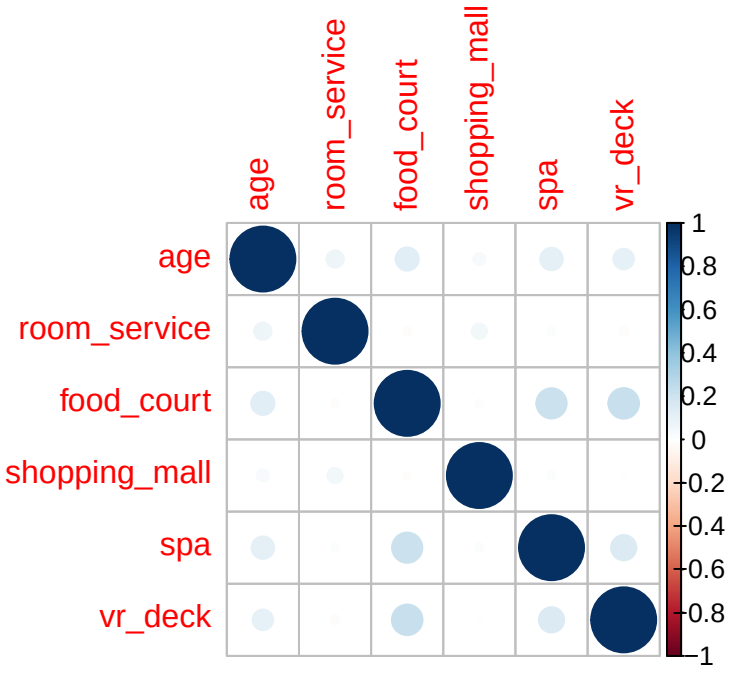
Passengers are more likely to not be transported if they a large amount of money on VR Deck.

Exercise 6

Make a correlogram of the continuous variables in the dataset. What do you observe?

```
cor_matrix <- spaceship_data %>%
  mutate(across(c(age, room_service, food_court, shopping_mall, spa, vr_deck), as.numeric)
  select(age, room_service, food_court, shopping_mall, spa, vr_deck) %>%
  cor(use = "complete.obs")

corrplot(cor_matrix, method = "circle")
```



None of the continuous variables exhibit a strong correlation in this dataset. However, spa and food court, as well as vr deck and food court, have slightly higher correlations than the other variables.

Exercise 7

Use binning to divide the feature `age` into six groups: ages 0-12, 13-17, 18-25, 26-30, 31-50, and 51+.

```
spaceship_data <- spaceship_data %>%
  mutate(age_binned = cut(as.numeric(age), breaks = c(0, 12, 17, 25, 30, 50, Inf),
                          include.lowest = T,
                          labels = c("0-12", "13-17", "18-25", "26-30", "31-50", "51+"),
                          right = T))
```

Exercise 8

For the expenditure variables, do the following:

- Create a new feature that consists of the total expenditure across all five amenities;

- Create a binary feature to flag passengers who did not spend anything (a total expenditure of 0);
- Log-transform the total expenditure to reduce skew.

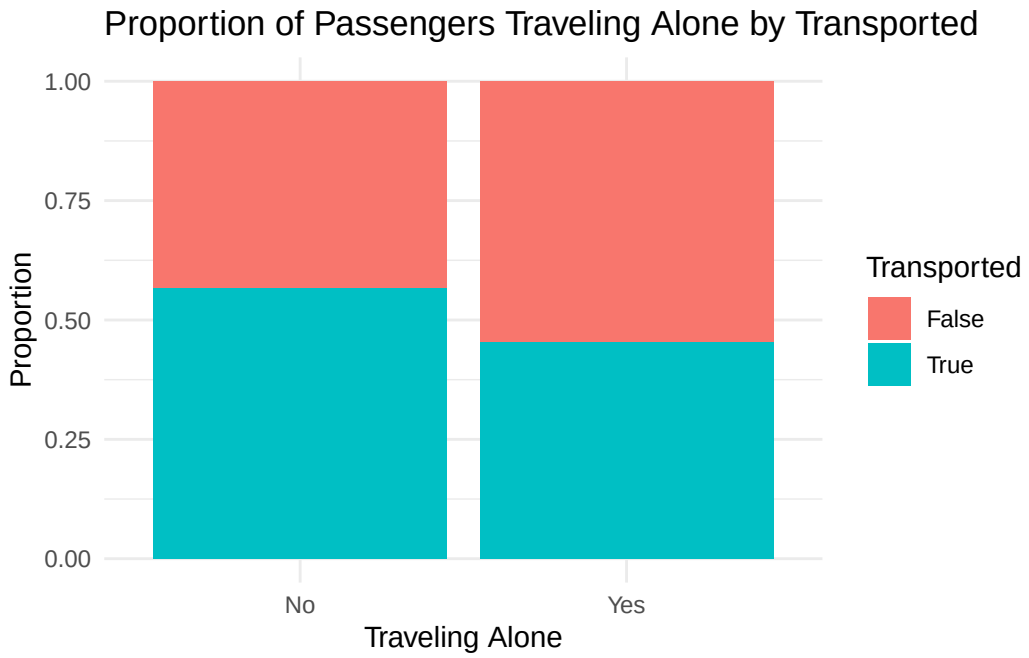
```
spaceship_data <- spaceship_data %>%
  mutate(total_expenditure = room_service + food_court + shopping_mall + spa + vr_deck,
    no_expenditure = ifelse(total_expenditure == 0, 1, 0), # 1: no expenditure
    log_total_expenditures = log1p(total_expenditure))
```

Exercise 9

Using the `passenger_id` column, create a new binary-coded feature that represents whether a passenger was traveling alone or not. Make a proportion stacked bar chart of this feature and `transported`. What do you observe?

```
spaceship_data <- spaceship_data %>%
  separate(passenger_id, into = c("group_id", "individual_id"), sep = "_") %>%
  group_by(group_id) %>%
  mutate(traveling_alone = ifelse(n() == 1, "Yes", "No")) %>%
  ungroup()

ggplot(spaceship_data, aes(x = as.factor(traveling_alone), fill = transported)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Passengers Traveling Alone by Transported",
       x = "Traveling Alone",
       y = "Proportion",
       fill = "Transported") +
  theme_minimal()
```



Passengers are less likely to be transported if they are traveling alone.

Exercise 10

Using the `cabin` variable, extract:

1. Cabin deck (A, B, C, D, E, F, G, or T);
2. Cabin number (0 to 2000);
3. Cabin side (P or S).

```
spaceship_data <- spaceship_data %>%
  separate(cabin, into = c("cabin_deck", "cabin_number", "cabin_side"), sep = "/", convert
```

Then do the following:

- Drop any observations with a cabin deck of T;
- Bin cabin number into groups of 300 (for example, 0 - 300, 301 - 600, 601- 900, etc.).

```
spaceship_data <- spaceship_data %>%
  filter(cabin_deck != "T") %>%
  mutate(cabin_number_binned = cut(cabin_number, breaks = c(0, seq(301, 2000, by = 300), I
      right = T,  # Upper bound inclusive
```

```
        include.lowest = T, # Include 0
        labels = c("0-300", paste(seq(301, 1700, by = 300), "-", seq(600, 2000, by = 300), s
```

Exercise 11

Create a new data frame (or tibble) that retains the following features:

1. `home_planet`
2. cabin deck
3. cabin number (binned)
4. cabin side
5. `age` (binned)
6. total expenditures (log-transformed)
7. `cryo_sleep`
8. `destination`
9. whether the passenger was traveling alone (call this `solo`)

To those features, do the following:

- One-hot encode all categorical features
- Center and scale all continuous features

```
new_spaceship_data <- spaceship_data %>%
  mutate(solo = traveling_alone) %>%
  select(home_planet, cabin_deck, cabin_number_binned, cabin_side, age_binned, log_total_e

spaceship_recipe <- recipe(~ ., new_spaceship_data) %>%
  step_dummy(all_nominal_predictors(), one_hot = T) %>%
  step_center(all_numeric_predictors()) %>%
  step_scale(all_numeric_predictors())

spaceship_prep_bake <- spaceship_recipe %>%
  prep() %>%
  bake(new_data = NULL)
```

```
New names:
* `cabin_number_binned_X1801...2000` -> `cabin_number_binned_X1801`
```

```
head(spaceship_prep_bake, 5)
```

```
# A tibble: 5 x 33
  log_total_expenditures home_planet_Earth home_planet_Europa home_planet_Mars
                   <dbl>             <dbl>              <dbl>            <dbl>
1                 -1.16             -1.11               1.76           -0.504
2                  0.627             0.898             -0.567          -0.504
3                  1.34             -1.11               1.76           -0.504
4                  1.15             -1.11               1.76           -0.504
5                  0.734             0.898             -0.567          -0.504
# i 29 more variables: cabin_deck_A <dbl>, cabin_deck_B <dbl>,
#   cabin_deck_C <dbl>, cabin_deck_D <dbl>, cabin_deck_E <dbl>,
#   cabin_deck_F <dbl>, cabin_deck_G <dbl>, cabin_number_binned_X0.300 <dbl>,
#   cabin_number_binned_X301.600 <dbl>, cabin_number_binned_X601.900 <dbl>,
#   cabin_number_binned_X901.1200 <dbl>, cabin_number_binned_X1201.1500 <dbl>,
#   cabin_number_binned_X1501.1800 <dbl>, cabin_number_binned_X1801 <dbl>,
#   cabin_side_P <dbl>, cabin_side_S <dbl>, age_binned_X0.12 <dbl>, ...
```

Exercise 12

Write up your analyses thus far in one or two paragraphs. Describe what you have learned about the passengers on the Spaceship Titanic. Describe the relationships you observed between variables. Which features do you think may be the best predictors of transported status? Why or why not?

By thorough research on the Spaceship Titanic dataset, I find out that home_planet, cryo_sleep, destination,VIP status all have a impact on the chance of being transported. In addition, age seems to be an important predictor because passengers under 20 are more likely to be transported. room_service, food_court, shopping_mall, spa, and vr_deck are also very important variables, as it appears that passengers who spend more money on these services are less likely to be transported. This might suggest that socio-economic status could significantly influence survival. Additionally, these variables do not have strong correlations with each other, which makes them useful for modeling. Whether individuals traveled alone also seems to affect the likelihood of being transported. In conclusion, I would say most variables excepted name and passenger_id are good predictors of transported status.

Exercises for 234 Students

Exercise 13

Split the dataset into training and testing. Use random sampling. Make sure that 80% of observations are in the training set and the remaining 20% in the testing set.

Exercise 14

Using k-fold cross-validation with k of 5, tune two models:

1. A random forest;
2. An elastic net.

Exercise 15

Select your best **random forest** model and use it to predict your testing set. Present the following:

- Testing accuracy;
- Testing ROC curve (and area under the ROC curve);
- Confusion matrix;
- Variable importance plot.

Exercise 16

Write up your conclusions in one to two paragraphs. Answer the following: How did your models do? Are you happy with their performance? Is there another model (besides these two) that you would be interested in trying? Which features ended up being the most important in terms of predicting whether or not a passenger would be transported?