Homework 2

Due: March 9, 2018 @ 11:59pm

Instructions

Obtain the Github repository you will use to complete homework 2 that contains a starter RMarkdown file named `homework_2.Rmd`, which you will use to do your work and write-up when completing the questions below. Remember to fill in your name at the top of the RMarkdown document and be sure to save, commit, and push (upload) frequently to Github so that you have incremental snapshots of your work. When you're done, follow the How to submit section below to setup a Pull Request, which will be used for feedback.

- Remember that the point of us using RMarkdown documents is to combine code and writeups! Each block of R code should have some sort of explanation or justification using full sentences.
- **Your grade will take into account your code, your explanations, and whether your document looks nice when "knitted" to HTML or PDF.**

Overview

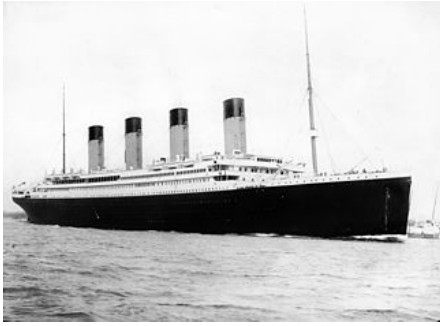

Figure 1: A photograph of the *Titanic* leaving Southampton on April 10, 1912.

For this homework assignment, you will be exploring a dataset about the passengers on the *Titanic*, the British passenger liner that crashed into an iceberg during its maiden voyage and sank early in the morning on April 15, 1912. The tragedy stands out as one of the deadliest commercial maritime disasters during peacetime in history. More than half of the passengers and crew died, due in large part to poor safety standards, such as not having enough lifeboats or not ensuring all lifeboats were filled to capacity during evacuation.

This dataset presents the most up-to-date knowledge about the passengers that were on the *Titanic*, including whether or not they survived. This dataset is frequently used to introduce using machine learning techniques that take multiple inputs and use them to predict an outcome, in this case whether a passenger is likely to have survived. While we won't be using a machine learning model in this assignment, there is still a lot of information that can be learned by exploring the dataset using the `tidyverse` suite.

The dataset is included in your Github starter code repository.

About the dataset

The following are the variable (column) descriptions for the dataset:[1]

Variable	Description
`pclass`	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
`survival`	Survival (0 = No; 1 = Yes)
`name`	Name
`sex`	Sex
`age`	Age
`sibsp`	Number of Siblings/Spouses Aboard
`parch`	Number of Parents/Children Aboard
`ticket`	Ticket Number
`fare`	Passenger Fare (British pound)
`cabin`	Cabin
`embarked`	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
`boat`	Lifeboat
`body`	Body Identification Number
`home.dest`	Home/Destination

Also note that the following definitions were used for `sibsp` and `parch`:

Label	Definition
Sibling	Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic
Spouse	Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)
Parent	Mother or Father of Passenger Aboard Titanic
Child	Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Questions

1. When reading in the dataset using `read_csv(file = "titanic_dataset.csv")`, several of the columns are converted into inconvenient data types. Fix this so that your later analysis does not run into problems. Use the `col_types = cols()` argument within `read_csv()`, see this section of *R for Data Science* for a review, to change the data type defaults for the following columns:
   - Convert `survived` to the logical data type
   - Convert `pclass` to the character data type
   - Convert `sibsp` to the character data type
   - Convert `parch` to the character data type
2. Compute how many known passengers were on the Titanic. *Do not just print the table, use a function to count the passengers.*
3. A famous directive for evacuating the *Titanic* was "women and children first". Use your `dplyr` functions to verify the first part of this statement by counting the number of men and women that survived and that died. Then, using those counts, calculate the fraction of women that survived,

[1] http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.html

$$\frac{\text{Number of female survivors}}{\text{Total number of female passengers}}$$

and the fraction of men that survived,

$$\frac{\text{Number of male survivors}}{\text{Total number of male passengers}}$$

Do your computations support the idea that women were more likely to survive? Why or why not?

4. Verify the second part of the "women and children first" directive. This will not be as straightforward as it was in the previous question, as the dataset only contains people's ages, which can take on many values. By default, there are no columns with labels of *child* or *adult*, so you will need to create your own.

   Create a new column named `child_or_adult` that uses the age data to label each passenger. For our purposes, we want to label anyone aged 0–9 as a child and anyone age 10 and up are as adults. If the age cell is blank (NA) for a passenger, also label them as an "adult". Assign this updated dataset to the variable `titanic_age_groups`.

   Hint: You will need to use the `ifelse()` function to complete this task. An example usage of `ifelse()` is the following:

```
titanic %>%
  mutate(cheap_or_expensive = ifelse(test = fare < 15,
                                     yes = "cheap ticket",
                                     no = "not cheap"))
```

   To handle blank entries, you will also need to use `is.na()` somewhere inside your `ifelse()` test.

5. Using the `titanic_age_groups` dataset you created in the previous question, count the number of children that survived and the number that did not. Do your computations support the idea that children were also more likely to survive? Why or why not?

6. A passenger's age group and sex are not the only predictors of survival. For example, social standing and wealth can play a factor in survival. One of the parameters within this dataset acts as a proxy for distinguishing between the upper and lower classes. Which parameter is it? How do you know?

7. Group your dataset by `sex` and the variable you determined in question 6 and count the number that survived and the number that did not. Create a bar chart that summarizes the data, where `survived` is along the horizontal axis and the passenger counts are along the vertical axis. Use the bar chart `fill =` aesthetic to break the bar charts down by your variable from question 6. Additionally, facet over the `sex` variable. Interpret this visualization and describe any survival patterns that you notice.

8. Create two visualizations:

   - The first visualization should be a bar chart displaying the fraction of the passengers that survived for different values of `parch`,

     $$\frac{\text{For a given parch, the number of survivors}}{\text{Total number of passengers}}.$$

     Doing this requires grouping your data properly, counting the number of passengers in each grouping, and then dividing this by the total number of passengers on the ship.

   - The second visualization should be a bar chart displaying the fraction of the passengers that survived for different values of `sibsp`,

     $$\frac{\text{For a given sibsp, the number of survivors}}{\text{Total number of passengers}}.$$

Like above, doing this requires grouping your data properly, counting the number of passengers in each grouping, and then dividing this by the total number of passengers on the ship.

Interpret the patterns that you see in the visualizations.

9. Based on your analysis, write a list of the factors that affected the chances of survival for each passenger. You should be able to identify 4 different attributes that had a noticeable impact on survival. Justify each attribute that you list by referencing back to a table or visualization you created in a previous question.

How to submit

When you are ready to submit, be sure to save, commit, and push your final result so that everything is synchronized to Github. Then, navigate to **your copy** of the Github repository you used for this assignment. You should see your repository, along with the updated files that you just synchronized to Github. Confirm that your files are up-to-date, and then do the following steps:

1. Click the *Pull Requests* tab near the top of the page.
2. Click the green button that says "New pull request".
3. Click the dropdown menu button labeled "base:", and select the option `starting`.
4. Confirm that the dropdown menu button labeled "compare:" is set to `master`.
5. Click the green button that says "Create pull request".
6. Give the *pull request* the following title: `Submission: Homework 2, FirstName LastName`, replacing `FirstName` and `LastName` with your actual first and last name.
7. In the messagebox, write: `My homework submission is ready for grading @shuaibm @jkglasbrenner`.
8. Click "Create pull request" to lock in your submission.

Cheatsheets

You are encouraged to review and keep the following cheatsheets handy while working on this assignment:

- RStudio cheatsheet
- RMarkdown cheatsheet
- RMarkdown reference
- `ggplot2` cheatsheet
- Data transformation cheatsheet
- Data import cheatsheet