Lab 5: Telling a story about survival on the Titanic

Create a short reproducible report answering the questions below, The report should be less than 3 pages, including all figures, and should be submitted as both PDF and Rmd formats. You do not need to show your code in the PDF report.

This lab is due at 5pm on Friday, October 20th. You should submit your assignment, in the form of both a knitted RMarkdown PDF as well as the .Rmd file that created the PDF, by uploading them to your personal Google Drive folder that is shared with the TA and the instructor. While you may collaborate with other students on this assignment, you must write up your own code and answers to the questions. Absolutely no cutting and pasting of any portion of the answers. This assignment, like the others, will be worth 50 points.

Introduction to the Titanic dataset

In this exercise, we are going to look at a dataset describing characteristics (including survival) of passengers on the Titanic, which sunk in the North Atlantic Ocean in 1912.

Getting started

Let's load the data into our current R session, and look at the variables available in the dataset:

```
library("Hmisc")
getHdata(titanic)
head(titanic)
```

Exercise 1 **Understand your data** For this analysis, we are going to focus primarily on the impact of three predictor variables on survival: economic status (`pclass`), `sex`, and `age`. Examine your data carefully using some univariate plots and/or summaries of the variables to understand what the distributions look like.

Exercise 2 **Deal with missing data** There are a lot of missing data in the age variable. In real data analysis problems, missing data is a common and pesky problem. It can especially be difficult to deal with when the missingness is not "random", i.e. certain factors (whether they are variables you measure or not) can predict whether the data will be missing or not. For example, do you think that older or younger people might be more or less likely to be missing their age from this dataset? What other variables might determine whether we have age or not? One way to start to look at this is to create a new factor variable that indicates whether 'age' is missing or not for each observation. Then we can create some simple tables to assess missingness across different groups. Try these types of tabulations out and determine whether you think missing age is predicable based on some of the other data in our dataset.

```
titanic$age_mis <- factor(is.na(titanic$age))
mosaic::tally(~sex|age_mis, data=titanic)
```

For now, to make the rest of the lab easier to work through, we are going to ignore the observations that are missing age. This is rarely a great assumption to make in practice, especially if the missingness may be associated with other factors. So when we interpret our results, we will need to remember that our dataset may no longer be representative of the entire population of travelers on the Titanic. Run the following code to remove anyone missing age from our dataset.

```
titanic1 <- dplyr::filter(titanic, age_mis==FALSE)
```

Exercise 3 **Make some hypotheses** *Before making any multivariate plots*, discuss with your teammates your hypotheses about what relationships might exist between these variables. Make a short list, including directions of possible relationships and possible interactions. Sketch out a few graphics that you want to make.

Exercise 4 **Look at your data** Create a few exploratory graphics and/or tables that illustrate the relationships between these variables and survivial. (Hint: try adapting graphing code from the logistic regression lecture. Try using facets or colors to highlight important comparisons.)

Exercise 5 **Design and fit a model** Using the plots as your guide, write down a model that you'd like to fit to use to describe how this data predicts the outcome of survival. Then fit that model.

Exercise 6 **Examine your model performance** Now that you've fit a model, you can calculate, for each individual, an estimated probability of survival, using code similar to that below (Note: adding 'type='response'' ensures that if you fit a logistic regression model, the 'predict' function returns you predicted probabilities instead of predicted log-odds.):

```
titanic1$preds <- predict(fm1, type = "response")
```

Using these predicted probabilities, try to determine where your model performs worst. Start by figuring out a metric that you could use to measure a poor prediction for a particular observation. Then, summarize your data in a way that shows you what type of individuals you make the worst predictions about. Why did you make poor predictions for these subsets of people? What other data would you like to have to help you make even better predictions?