Lab 10

Titanic Survival Prediction

This lab will return to the titanic dataset to predict survival of a passenger. This dataset is obtained from the `earth` package in R.

The data frame has 1046 observations on 6 variables.

- pclass passenger class, unordered factor: 1st 2nd 3rd
- survived integer: 0 or 1
- sex unordered factor: male female
- age age in years, min 0.167 max 80.0
- sibsp number of siblings or spouses aboard, integer: 0…8
- parch number of parents or children aboard, integer: 0…6

```
set.seed(11142024)
library(tidyverse)
library(earth)
data("etitanic")
titanic <- etitanic |>
  mutate(survived_factor = factor(survived))
glimpse(titanic)
```

```
Rows: 1,046
Columns: 7
$ pclass          <fct> 1st, 1st, 1st, 1st, 1st, 1st, 1st, 1st, 1st, 1st, 1st,~
$ survived        <int> 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, ~
$ sex             <fct> female, male, female, male, female, male, female, male~
$ age             <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0000, 48.0000, 63~
$ sibsp           <int> 0, 1, 1, 1, 1, 0, 1, 0, 2, 0, 1, 1, 0, 0, 0, 0, 0, 0, ~
$ parch           <int> 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, ~
$ survived_factor <fct> 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, ~
```

Question 1 (2 points)

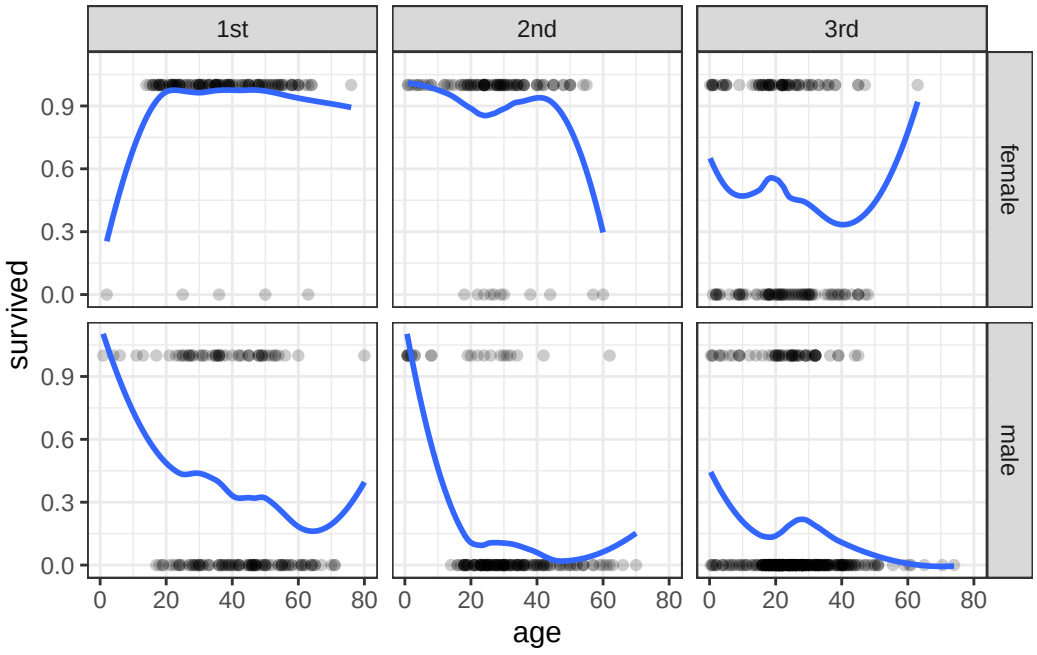
What factors in the dataset do you think will influence whether a passenger survives? How do you expect that factors to change survival outcomes?

Passenger class, sex, and age will likely be important. I'd expect an interaction between those factors.

Question 2 (6 points)

Regardless of your response to the first question, create figures to explore survival as a function of `age`, `sex`, and `pclass`. As always, include informative titles, axes, and legends.

```
titanic |>
  ggplot(aes(y = survived, x = age)) +
  geom_point(alpha = .2) +
  geom_smooth(method = 'loess', formula = 'y~x', se = F) +
  facet_grid(sex ~ pclass) +
  theme_bw()
```



Question 3 (2 points)

Construct a training set and a test set. If you plan to do model tuning, also create a validation set.

```
titanic <- titanic |>
  mutate(passenger_id = 1:n())

train_ids <- sample(1:nrow(titanic), ceiling(nrow(titanic) * .7))
train_titanic <- titanic |>
  filter(passenger_id %in% train_ids)
test_titanic <- titanic |>
  filter(!passenger_id %in% train_ids)
```

Question 4 (4 points)

Use a logistic regression model to predict passenger survival. Summarize the model outcome using classification error (% of incorrect predictions on the test set).

```
log_reg <- glm(survived_factor ~ age * pclass * sex, family = binomial(link = 'logit'), data
summary(log_reg)
```

```
Call:
glm(formula = survived_factor ~ age * pclass * sex, family = binomial(link = "logit"),
    data = train_titanic)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.06346    1.29649   1.592   0.1115
age                    0.03103    0.03749   0.828   0.4079
pclass2nd              0.68292    1.58541   0.431   0.6666
pclass3rd             -2.10587    1.35423  -1.555   0.1199
sexmale               -0.72702    1.46911  -0.495   0.6207
age:pclass2nd         -0.05854    0.04629  -1.265   0.2059
age:pclass3rd         -0.03517    0.04092  -0.860   0.3901
age:sexmale           -0.07996    0.04113  -1.944   0.0519 .
pclass2nd:sexmale     -0.24372    1.93039  -0.126   0.8995
pclass3rd:sexmale     -0.31654    1.58157  -0.200   0.8414
age:pclass2nd:sexmale -0.04978    0.06213  -0.801   0.4230
age:pclass3rd:sexmale  0.06506    0.04710   1.381   0.1672
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 994.04  on 732  degrees of freedom
Residual deviance: 634.36  on 721  degrees of freedom
AIC: 658.36

Number of Fisher Scoring iterations: 6
```

```
ce <- 1 - mean(test_titanic$survived == round(predict(log_reg, newdata = test_titanic, type =
```

The classification error for this logistic regression model is 21.1%

Question 5 (4 points)

Use a tree-based model to predict passenger survival. Summarize the model outcome using classification error (% of incorrect predictions on the test set).

```
library(randomForest)
rf <- randomForest(survived_factor ~ age + pclass + sex + sibsp + parch, family = binomial(li
print(rf)
```

```
Call:
 randomForest(formula = survived_factor ~ age + pclass + sex +      sibsp + parch, data = tra
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 20.74%
Confusion matrix:
    0   1 class.error
0 383  47   0.1093023
1 105 198   0.3465347
```

```
ce_rf <- 1 - mean(test_titanic$survived == predict(rf, newdata = test_titanic, type = 'respon
```

The random forest algorithm predict a classification error of 21.1%

Question 6 (2 points)

Do your model outcomes match your intuition and data visualizations? Why or why not?

Yes, I had pretty good understanding of the factors that would influence survival - thanks James Cameron.