Homework 3

Jules Merigot (8488256)

October 19, 2022

Classification

Question 1

```
# setting the seed
set.seed(8488)

titanic_split <- initial_split(titanic_data, prop=0.60, strata=survived)

titanic_train <- training(titanic_split)
titanic_test <- testing(titanic_split)
```

For splitting the data, I chose a proportion of 0.60 because it allows for more training data, while retaining enough data to be tested since there is a limited amount of observations. The training data has 534 observations while the testing data has 357 observations.

```
# missing value in the training data
head(is.na(titanic_train))
```

```
##    passenger_id survived pclass  name   sex   age sib_sp parch ticket  fare
## 6         FALSE    FALSE  FALSE FALSE FALSE  TRUE  FALSE FALSE  FALSE FALSE
## 7         FALSE    FALSE  FALSE FALSE FALSE FALSE  FALSE FALSE  FALSE FALSE
## 8         FALSE    FALSE  FALSE FALSE FALSE FALSE  FALSE FALSE  FALSE FALSE
## 15        FALSE    FALSE  FALSE FALSE FALSE FALSE  FALSE FALSE  FALSE FALSE
## 17        FALSE    FALSE  FALSE FALSE FALSE FALSE  FALSE FALSE  FALSE FALSE
## 21        FALSE    FALSE  FALSE FALSE FALSE FALSE  FALSE FALSE  FALSE FALSE
##    cabin embarked
## 6   TRUE    FALSE
## 7  FALSE    FALSE
## 8   TRUE    FALSE
## 15  TRUE    FALSE
## 17  TRUE    FALSE
## 21  TRUE    FALSE
```

```
# the number of missing values in the training data
sum(is.na(titanic_train))
```

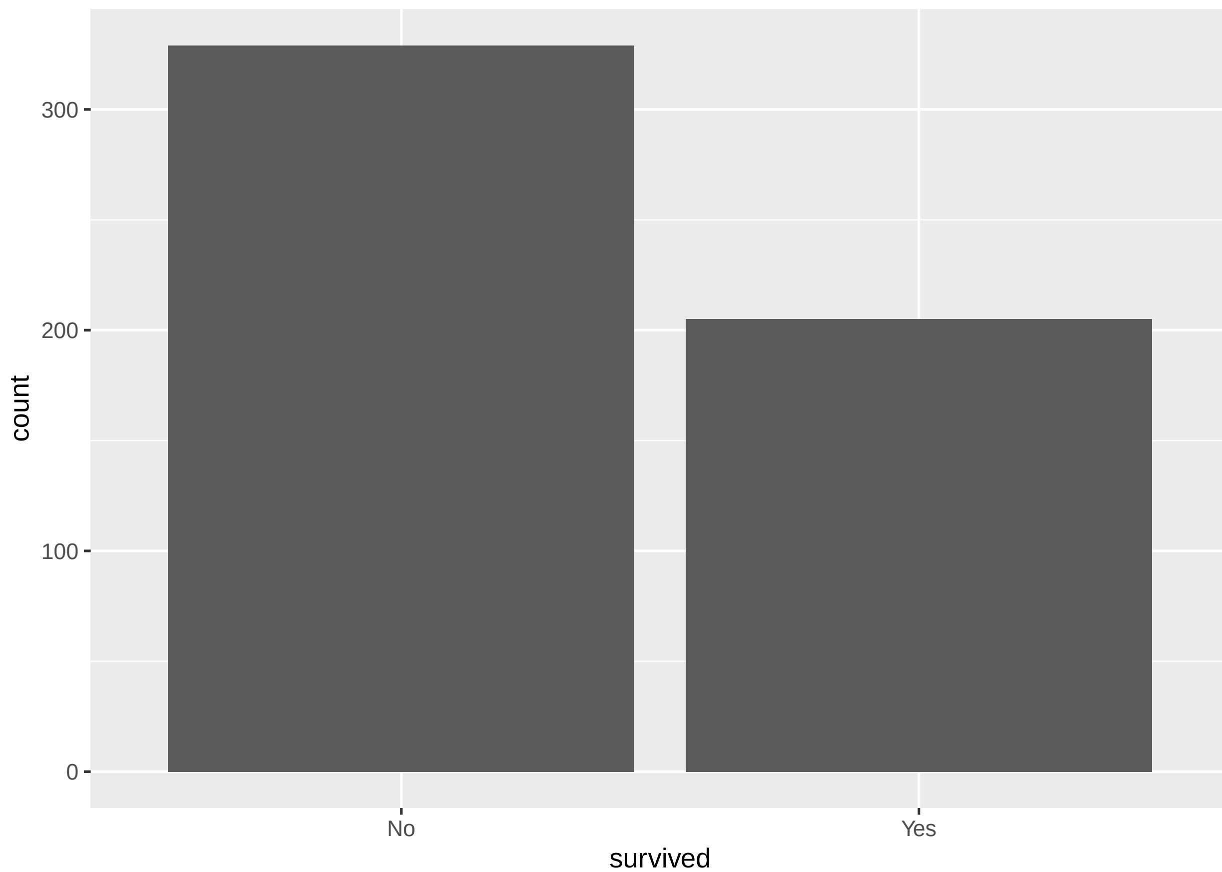
```
## [1] 522
```

There is a good amount of missing data in the training data, 522 missing data values to be exact, especially for the *age* variable, as can be seen using the code above.

We want to use stratified sampling for this data because since we have less observations than the abalone dataset for example, stratified sampling allows for more precision on a smaller dataset, and thus a more precise sample in this case.

Question 2

```
titanic_train %>%
  ggplot(aes(x = survived)) +
  geom_bar()
```



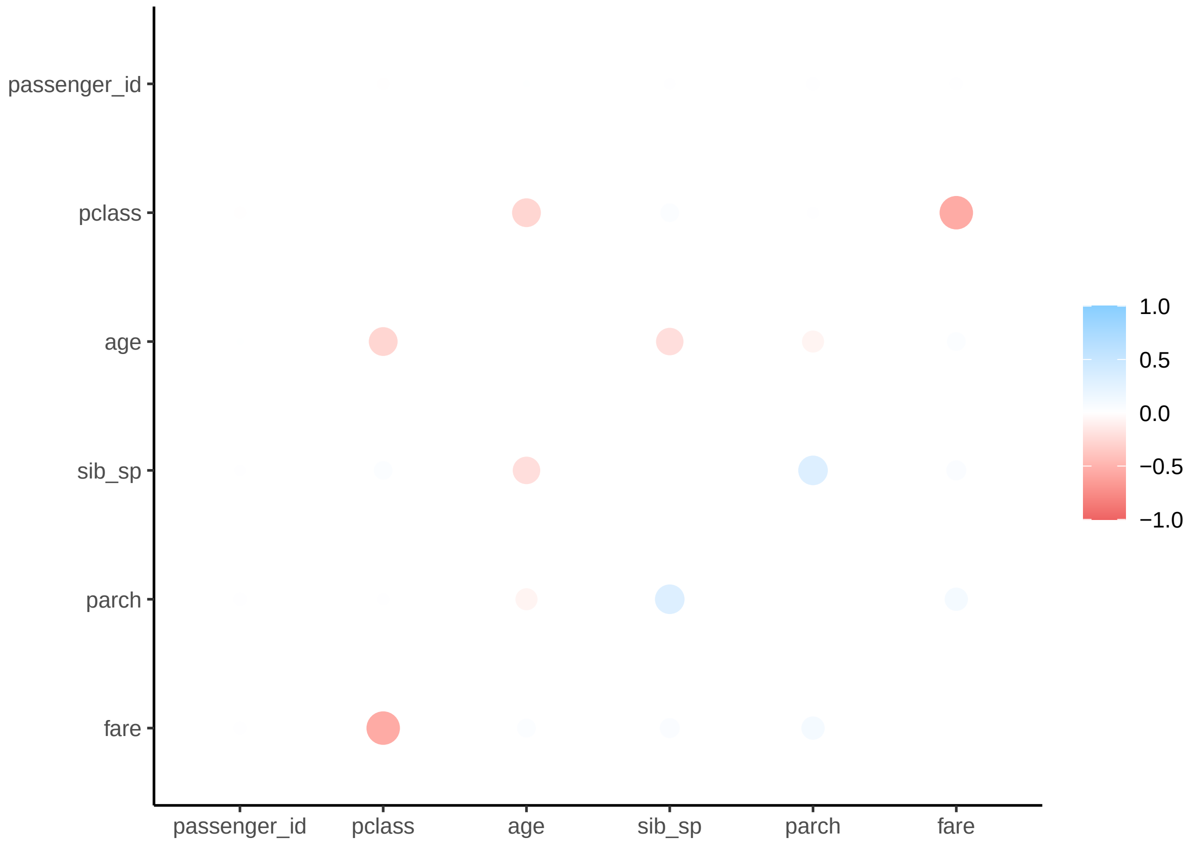
Using the above visualization of the distribution of the outcome variable *survived*, we can see that less people survived than people that perished on the Titanic. More than 300 (known) passengers lost their lives, while only a little more than 200 passengers survived.

Question 3

```
cor_titanic <- titanic_train %>%
  select(where(is.numeric)) %>%
  correlate()
```

```
## Correlation computed with
## * Method: 'pearson'
## * Missing treated using: 'pairwise.complete.obs'
```

```
rplot(cor_titanic)
```



After making the correlation matrix above, there are some clear patterns that emerge, such as most variables being slightly negatively correlated with others, with some exceptions. *parch* and *sib_sp* have a positive correlation, which means that the # of siblings/spouses of a certain passenger is positively correlated with the # of children/parents of that passenger, which makes sense. Additionally, *fare* and *pclass* are negatively correlated, which indicates that a passenger's fare is negatively correlated with the class of their ticket. This also makes sense, since as passenger's ticket class number decreases from third to first (which is technically increasing), their fare will increase in price.

Question 4

```
titanic_recipe <- recipe(survived ~ pclass + sex + age + sib_sp + parch + fare,
                         data=titanic_data) %>%
  step_impute_linear(age, impute_with = imp_vars(all_predictors())) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ sex_male:fare + age:fare)

titanic_recipe %>% prep() %>% juice()
```

```
## # A tibble: 891 x 9
##    pclass   age sib_sp parch  fare survived sex_male sex_male_x_fare fare_x_age
##     <int> <dbl>  <int> <int> <dbl> <fct>       <dbl>           <dbl>      <dbl>
##  1      3  22        1     0  7.25 No              1            7.25      160.
##  2      1  38        1     0 71.3  Yes             0            0        2709.
##  3      3  26        0     0  7.92 Yes             0            0         206.
##  4      1  35        1     0 53.1  Yes             0            0        1858.
##  5      3  35        0     0  8.05 No              1            8.05      282.
##  6      3  28.3      0     0  8.46 No              1            8.46      239.
##  7      1  54        0     0 51.9  No              1           51.9      2801.
##  8      3   2        3     1 21.1  No              1           21.1        42.2
##  9      3  27        0     2 11.1  Yes             0            0         301.
## 10      2  14        1     0 30.1  Yes             0            0         421.
## # ... with 881 more rows
```

Question 5

```
log_reg <- logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")

log_wkflow <- workflow() %>%
  add_model(log_reg) %>%
  add_recipe(titanic_recipe)

log_fit <- fit(log_wkflow, titanic_train)

log_fit %>%
  tidy()
```

```
## # A tibble: 9 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)      5.81       0.947       6.14   8.38e-10
## 2 pclass          -1.31       0.217      -6.03   1.65e- 9
## 3 age             -0.0641     0.0151     -4.24   2.21e- 5
## 4 sib_sp          -0.756      0.169      -4.48   7.35e- 6
## 5 parch           -0.0159     0.160      -0.0996 9.21e- 1
## 6 fare             0.0320     0.0202      1.59   1.13e- 1
## 7 sex_male        -2.04       0.392      -5.21   1.85e- 7
## 8 sex_male_x_fare -0.0386     0.0170     -2.27   2.35e- 2
## 9 fare_x_age       0.0000619  0.000294    0.210  8.33e- 1
```

Question 6

```
lda_mod <- discrim_linear() %>%
  set_mode("classification") %>%
  set_engine("MASS")
```

```
lda_wkflow <- workflow() %>%
  add_model(lda_mod) %>%
  add_recipe(titanic_recipe)

lda_fit <- fit(lda_wkflow, titanic_train)
```

Question 7

```
qda_mod <- discrim_quad() %>%
  set_mode("classification") %>%
  set_engine("MASS")

qda_wkflow <- workflow() %>%
  add_model(qda_mod) %>%
  add_recipe(titanic_recipe)

qda_fit <- fit(qda_wkflow, titanic_train)
```

Question 8

```
nb_mod <- naive_Bayes() %>%
  set_mode("classification") %>%
  set_engine("klaR") %>%
  set_args(usekernel = FALSE)

nb_wkflow <- workflow() %>%
  add_model(nb_mod) %>%
  add_recipe(titanic_recipe)

nb_fit <- fit(nb_wkflow, titanic_train)
```

Question 9