```
options(warn = -1)
library(tidyverse)
library(ggthemes)
library(ggplot2)
library(tidymodels)
library(corrplot)
library(yardstick)
library(ISLR)
library(ISLR2)
library(discrim)
library(corrr)
tidymodels_prefer()
```

```
titanic <- read.csv(file="/Users/honchowayne/Desktop/titanic.csv")
titanic %>% head()
```

```
##   passenger_id survived pclass
## 1            1       No      3
## 2            2      Yes      1
## 3            3      Yes      3
## 4            4      Yes      1
## 5            5       No      3
## 6            6       No      3
##                                                  name    sex age sib_sp parch
## 1                             Braund, Mr. Owen Harris   male  22      1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38      1     0
## 3                              Heikkinen, Miss. Laina female  26      0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35      1     0
## 5                            Allen, Mr. William Henry   male  35      0     0
## 6                                    Moran, Mr. James   male  NA      0     0
##             ticket    fare cabin embarked
## 1        A/5 21171  7.2500  <NA>        S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250  <NA>        S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500  <NA>        S
## 6           330877  8.4583  <NA>        Q
```

```
titanic$survived <- as.factor(titanic$survived)
titanic$pclass <- as.factor(titanic$pclass)
titanic$sex <- as.factor(titanic$sex)
titanic$survived <- factor(titanic$survived, levels = c("Yes" , "No"))
titanic$survived %>% head()
```

```
## [1] No  Yes Yes Yes No  No
## Levels: Yes No
```

QUESTION1:

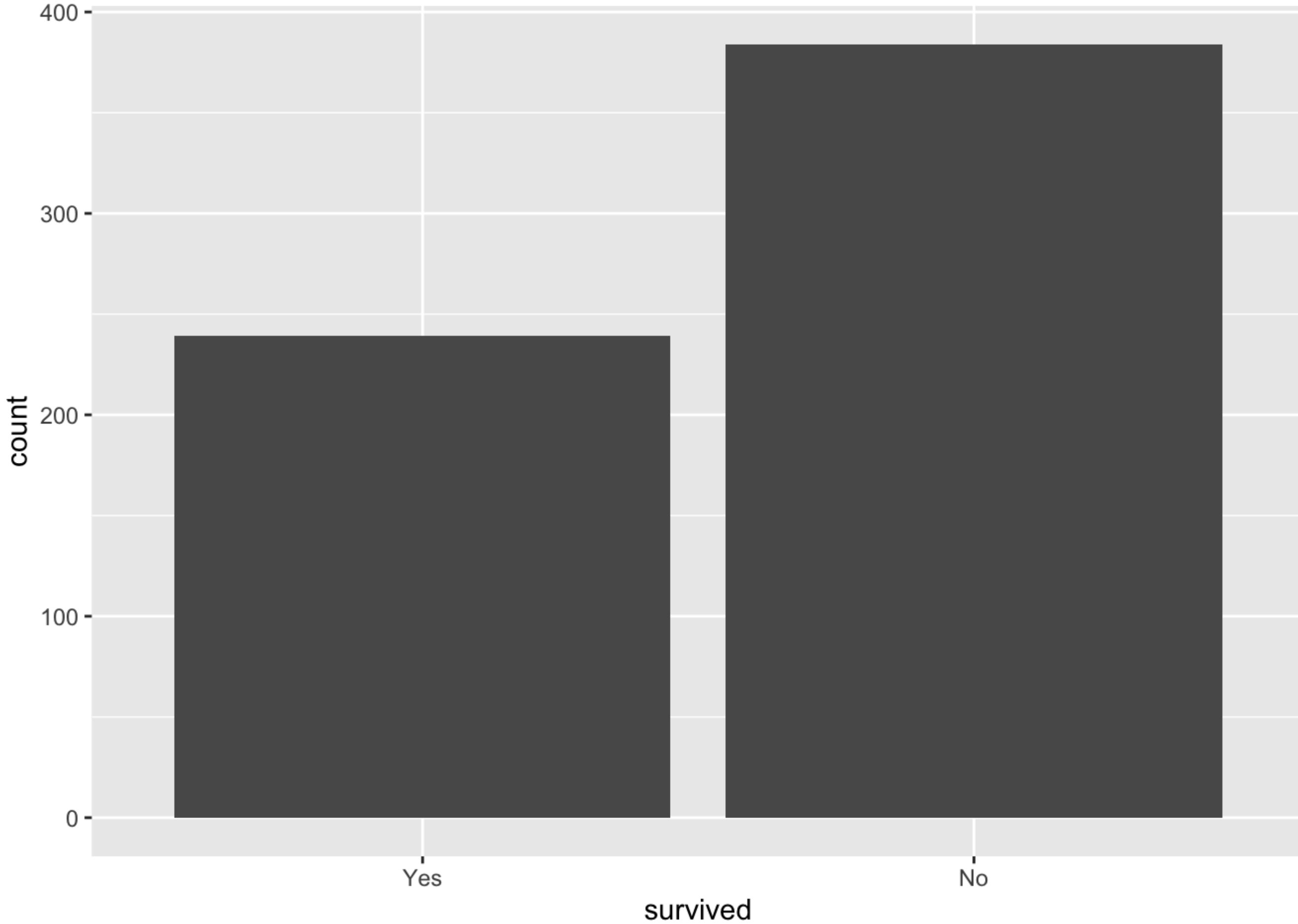
```
set.seed(3435)
titanic_split <- initial_split(titanic, prop = 0.70, strata = survived)
titanic_train <- training(titanic_split)
titanic_test <- testing(titanic_split)
titanic_train %>% head()
```

```
##    passenger_id survived pclass                              name    sex age
## 1             1       No      3           Braund, Mr. Owen Harris   male  22
## 5             5       No      3          Allen, Mr. William Henry   male  35
## 7             7       No      1           McCarthy, Mr. Timothy J   male  54
## 8             8       No      3    Palsson, Master. Gosta Leonard   male   2
## 14           14       No      3       Andersson, Mr. Anders Johan   male  39
## 15           15       No      3 Vestrom, Miss. Hulda Amanda Adolfina female  14
##    sib_sp parch    ticket    fare cabin embarked
## 1       1     0 A/5 21171  7.2500  <NA>        S
## 5       0     0    373450  8.0500  <NA>        S
## 7       0     0     17463 51.8625   E46        S
## 8       3     1    349909 21.0750  <NA>        S
## 14      1     5    347082 31.2750  <NA>        S
## 15      0     0    350406  7.8542  <NA>        S
```

```
#In short, stratified sampling ensures each subgroup within the population receives prop
er representation within the sample.
#some data in the "age" and "cabin" column are missing.
```

QUESTION2:

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




```
fct_count(titanic_train$survived)
```

```
## # A tibble: 2 × 2
##   f         n
##   <fct> <int>
## 1 Yes     239
## 2 No      384
```

```
# By inspection, the number of those who didn't survive in the accident is greater than
 those who survived (384 vs 239).
```

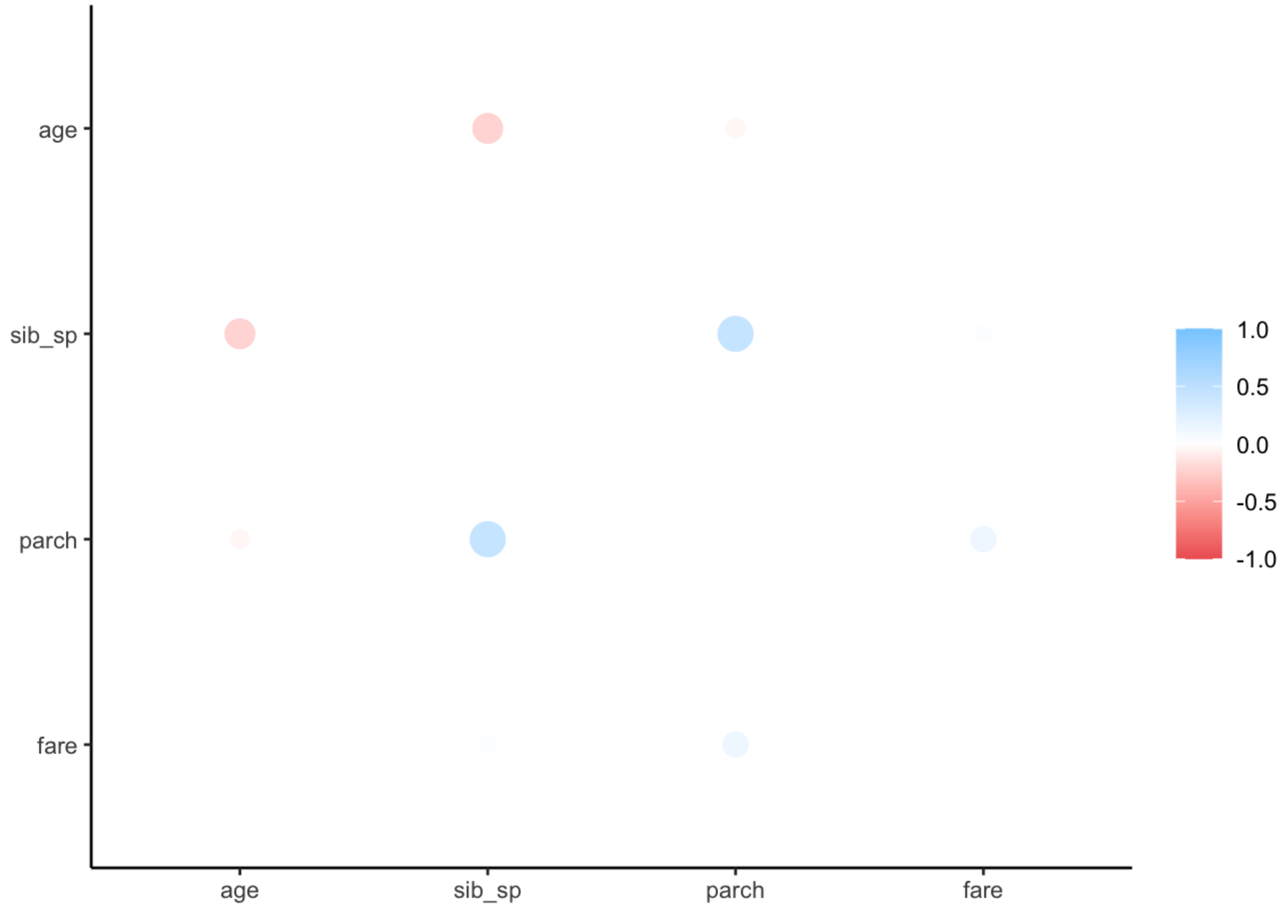
QUESTION3:

```
cor_titanic <- titanic_train %>% select(age, sib_sp, parch, fare) %>% correlate()
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
rplot(cor_titanic)
```

```
## Don't know how to automatically pick scale for object of type noquote. Defaulting to
continuous.
```




```
# Seems like the number of siblings / spouses aboard the Titanic and age have a weak neg
ative correlation
# the number of siblings / spouses aboard the Titanic and the number of parents / childr
en aboard the Titanic have a weak positive correlation.
```

QUESTION4:

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

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          6
##
## Operations:
##
## Linear regression imputation for age
## Dummy variables from all_nominal_predictors()
## Interactions with sex_male:fare
## Interactions with age:fare
```

QUSTION5:

```
# Specifying an Engine
log_reg_titanic <- logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")
```

```
log_titanicwkflow <- workflow() %>%
  add_model(log_reg_titanic) %>%
  add_recipe(titanic_recipe)
```

```
log_fit_titanic <- fit(log_titanicwkflow, titanic_train)
```

```
str(titanic_train)
```

```
## 'data.frame':    623 obs. of  12 variables:
##  $ passenger_id: int  1 5 7 8 14 15 17 19 25 28 ...
##  $ survived    : Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...
##  $ pclass      : Factor w/ 3 levels "1","2","3": 3 3 1 3 3 3 3 3 3 1 ...
##  $ name        : chr  "Braund, Mr. Owen Harris" "Allen, Mr. William Henry" "McCarthy,
Mr. Timothy J" "Palsson, Master. Gosta Leonard" ...
##  $ sex         : Factor w/ 2 levels "female","male": 2 2 2 2 2 1 2 1 1 2 ...
##  $ age         : num  22 35 54 2 39 14 2 31 8 19 ...
##  $ sib_sp      : int  1 0 0 3 1 0 4 1 3 3 ...
##  $ parch       : int  0 0 0 1 5 0 1 0 1 2 ...
##  $ ticket      : chr  "A/5 21171" "373450" "17463" "349909" ...
##  $ fare        : num  7.25 8.05 51.86 21.07 31.27 ...
##  $ cabin       : chr  NA NA "E46" NA ...
##  $ embarked    : chr  "S" "S" "S" "S" ...
```

```
log_fit_titanic %>%
  tidy()
```

```
## # A tibble: 10 × 5
##    term            estimate std.error statistic  p.value
##    <chr>              <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)     -4.40      0.683       -6.45  1.13e-10
##  2 age              0.0629    0.0136       4.62  3.77e- 6
##  3 sib_sp           0.437     0.132        3.30  9.52e- 4
##  4 parch            0.151     0.153        0.989 3.23e- 1
##  5 fare            -0.00116   0.0107      -0.108 9.14e- 1
##  6 pclass_X2        1.25      0.363        3.46  5.48e- 4
##  7 pclass_X3        2.44      0.382        6.39  1.62e-10
##  8 sex_male         2.15      0.303        7.09  1.32e-12
##  9 sex_male_x_fare  0.0139    0.00836      1.66  9.65e- 2
## 10 age_x_fare      -0.000360  0.000206    -1.75  8.03e- 2
```

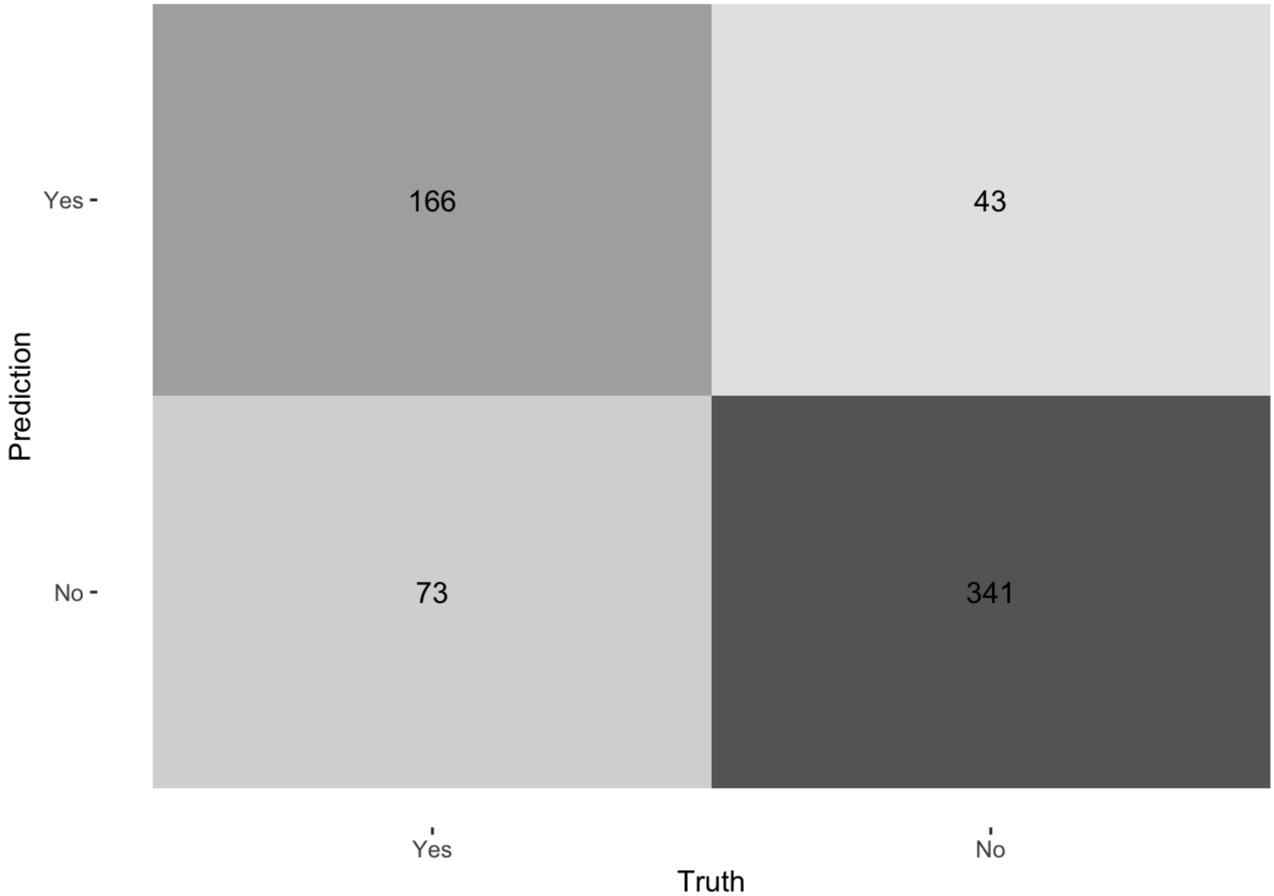
```
predict(log_fit_titanic, new_data = titanic_train, type = "prob")
```

```
## # A tibble: 623 × 2
##    .pred_Yes .pred_No
##        <dbl>    <dbl>
##  1    0.114     0.886
##  2    0.0835    0.916
##  3    0.310     0.690
##  4    0.115     0.885
##  5    0.0219    0.978
##  6    0.755     0.245
##  7    0.0711    0.929
##  8    0.449     0.551
##  9    0.519     0.481
## 10    0.108     0.892
## # … with 613 more rows
```

```
augment(log_fit_titanic, new_data = titanic_train) %>%
  conf_mat(truth = survived, estimate = .pred_class)
```

```
##           Truth
## Prediction Yes  No
##        Yes 166  43
##        No   73 341
```

```
augment(log_fit_titanic, new_data = titanic_train) %>%
  conf_mat(truth = survived, estimate = .pred_class) %>%
  autoplot(type = "heatmap")
```




```
log_reg_acc <- augment(log_fit_titanic, new_data = titanic_train) %>%
  accuracy(truth = survived, estimate = .pred_class)
log_reg_acc
```

```
## # A tibble: 1 × 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.814
```

QUESTION6:

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


lda_wkflow_titanic <- workflow() %>%
  add_model(lda_mod_titanic) %>%
  add_recipe(titanic_recipe)

lda_fit_titanic <- fit(lda_wkflow_titanic, titanic_train)
```

```
predict(lda_fit_titanic, new_data = titanic_train, type="prob")
```

```
## # A tibble: 623 × 2
##    .pred_Yes .pred_No
##        <dbl>    <dbl>
##  1    0.0762    0.924
##  2    0.0543    0.946
##  3    0.266     0.734
##  4    0.0872    0.913
##  5    0.0162    0.984
##  6    0.799     0.201
##  7    0.0592    0.941
##  8    0.528     0.472
##  9    0.614     0.386
## 10    0.167     0.833
## # … with 613 more rows
```

```
augment(lda_fit_titanic, new_data = titanic_train) %>%
  conf_mat(truth = survived, estimate = .pred_class)
```

```
##           Truth
## Prediction Yes  No
##        Yes 160  50
##        No   79 334
```

```
lda_acc<-augment(lda_fit_titanic, new_data = titanic_train) %>%
  accuracy(truth = survived, estimate = .pred_class)
lda_acc
```

```
## # A tibble: 1 × 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.793
```

QUESTION7:

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

qda_wkflow_titanic <- workflow() %>%
  add_model(qda_mod_titanic) %>%
  add_recipe(titanic_recipe)

qda_fit_titanic <- fit(qda_wkflow_titanic, titanic_train)
```

```
predict(qda_fit_titanic, new_data = titanic_train, type = "prob")
```

```
## # A tibble: 623 × 2
##      .pred_Yes  .pred_No
##          <dbl>     <dbl>
##  1 0.00971     0.990
##  2 0.00672     0.993
##  3 0.0962      0.904
##  4 0.000107    1.00
##  5 0.000000506 1.00
##  6 0.596       0.404
##  7 0.000000507 1.00
##  8 0.278       0.722
##  9 0.00100     0.999
## 10 1.00        0.0000288
## # … with 613 more rows
```

```
augment(qda_fit_titanic ,new_data = titanic_train) %>%
  conf_mat(truth = survived, estimate = .pred_class)
```

```
##           Truth
## Prediction Yes  No
##        Yes 128  27
##        No  111 357
```

```
qda_acc <- augment(qda_fit_titanic ,new_data = titanic_train) %>%
  accuracy(truth = survived, estimate = .pred_class)
qda_acc
```

```
## # A tibble: 1 × 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.778
```

QUESTION8:

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

nb_wkflow_titanic <- workflow() %>%
  add_model(nb_mod_titanic) %>%
  add_recipe(titanic_recipe)

nb_fit_titanic <- fit(nb_wkflow_titanic, titanic_train)
```

```
predict(nb_fit_titanic, new_data = titanic_train, type = "prob") %>% head()
```

```
## # A tibble: 6 × 2
##   .pred_Yes .pred_No
##       <dbl>    <dbl>
## 1 0.0269       0.973
## 2 0.0294       0.971
## 3 0.386        0.614
## 4 0.000228     1.00
## 5 0.0000220    1.00
## 6 0.521        0.479
```

```
nb_acc <- augment(nb_fit_titanic, new_data = titanic_train) %>%
  accuracy(truth = survived, estimate = .pred_class)
nb_acc
```

```
## # A tibble: 1 × 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.787
```

QUESTION9:

```
bind_cols(titanic_train$survived, predict(log_fit_titanic, new_data = titanic_train, typ
e = "prob"), predict(lda_fit_titanic, new_data = titanic_train, type="prob"), predict(qd
a_fit_titanic, new_data = titanic_train, type = "prob"), predict(nb_fit_titanic, new_dat
a = titanic_train, type = "prob"))
```

```
## New names:
## • `` -> `...1`
## • `.pred_Yes` -> `.pred_Yes...2`
## • `.pred_No` -> `.pred_No...3`
## • `.pred_Yes` -> `.pred_Yes...4`
## • `.pred_No` -> `.pred_No...5`
## • `.pred_Yes` -> `.pred_Yes...6`
## • `.pred_No` -> `.pred_No...7`
## • `.pred_Yes` -> `.pred_Yes...8`
## • `.pred_No` -> `.pred_No...9`
```

```
## # A tibble: 623 × 9
##    ...1  .pred_Yes...2 .pred_No...3 .pred_Yes...4 .pred_No...5 .pred_Yes...6
##    <fct>         <dbl>        <dbl>         <dbl>        <dbl>         <dbl>
##  1 No           0.114         0.886        0.0762        0.924   0.00971
##  2 No           0.0835        0.916        0.0543        0.946   0.00672
##  3 No           0.310         0.690        0.266         0.734   0.0962
##  4 No           0.115         0.885        0.0872        0.913   0.000107
##  5 No           0.0219        0.978        0.0162        0.984   0.000000506
##  6 No           0.755         0.245        0.799         0.201   0.596
##  7 No           0.0711        0.929        0.0592        0.941   0.000000507
##  8 No           0.449         0.551        0.528         0.472   0.278
##  9 No           0.519         0.481        0.614         0.386   0.00100
## 10 No           0.108         0.892        0.167         0.833   1.00
## # … with 613 more rows, and 3 more variables: .pred_No...7 <dbl>,
## #   .pred_Yes...8 <dbl>, .pred_No...9 <dbl>
```

```
accuracies <- c(log_reg_acc$.estimate, lda_acc$.estimate, nb_acc$.estimate, qda_acc$.est
imate)
models <- c("Logistic Regression", "LDA", "Naive Bayes", "QDA")
results <- tibble(accuracies = accuracies, models = models)
results %>% arrange(-accuracies)
```

```
## # A tibble: 4 × 2
##   accuracies models
##        <dbl> <chr>
## 1      0.814 Logistic Regression
## 2      0.793 LDA
## 3      0.787 Naive Bayes
## 4      0.778 QDA
```

```
# Based on the table listed above, Logistic Regression has the best performance on the t
raining data because its accuracy is the closest number to 1 among these models.
# Hence, I will apply Logistic Regression model because it has the most accurate result
 on the training data.
```

QUESTION10:

```
predict(log_fit_titanic,new_data = titanic_test, type = "prob")
```

```
## # A tibble: 268 × 2
##    .pred_Yes .pred_No
##        <dbl>    <dbl>
##  1    0.933    0.0671
##  2    0.924    0.0763
##  3    0.119    0.881
##  4    0.183    0.817
##  5    0.230    0.770
##  6    0.239    0.761
##  7    0.120    0.880
##  8    0.119    0.881
##  9    0.0456   0.954
## 10    0.174    0.826
## # … with 258 more rows
```

```
augment(log_fit_titanic, new_data = titanic_test) %>%
  conf_mat(truth = survived, estimate = .pred_class)
```

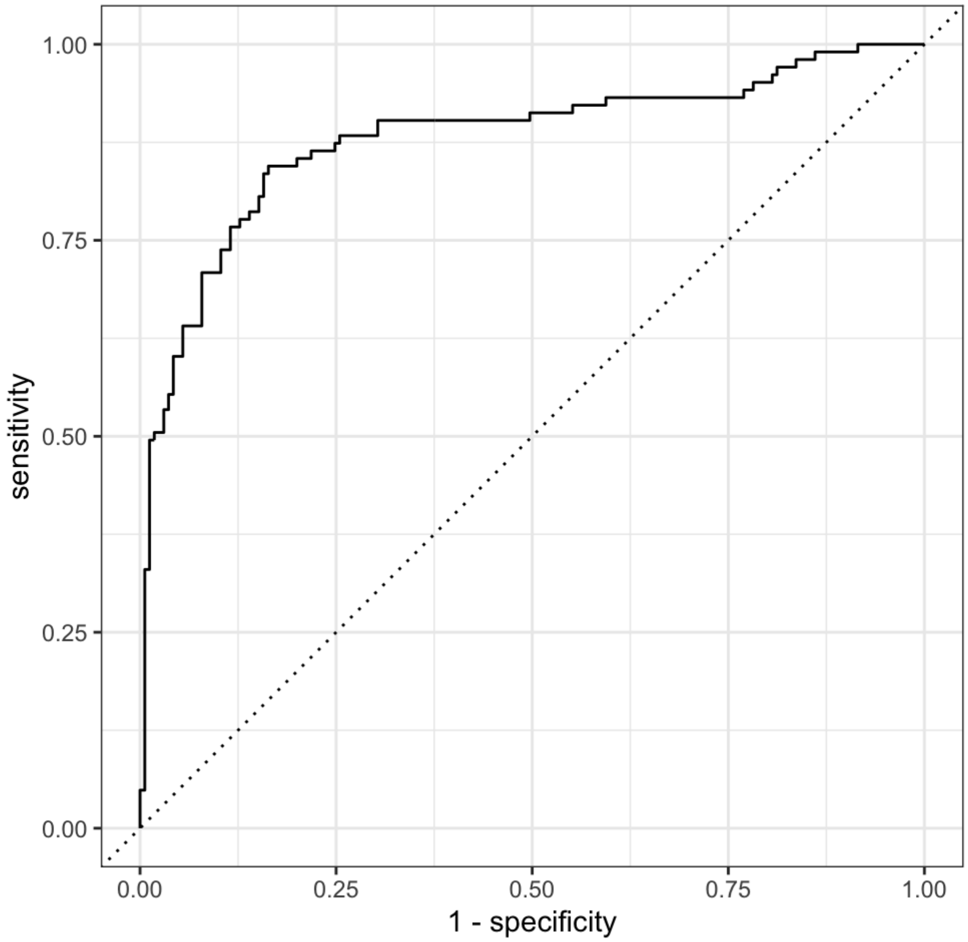
```
##           Truth
## Prediction Yes  No
##        Yes  75  17
##        No   28 148
```

```
#add two other metrics, sensitivity and specificity
multi_metric <- metric_set(accuracy, sensitivity, specificity)

augment(log_fit_titanic, new_data = titanic_test) %>%
  multi_metric(truth = survived, estimate = .pred_class)
```

```
## # A tibble: 3 × 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 accuracy    binary         0.832
## 2 sensitivity binary         0.728
## 3 specificity binary         0.897
```

```
# ROC curve
roc <- augment(log_fit_titanic, new_data = titanic_test) %>%
  roc_curve(survived, .pred_Yes) %>%
  autoplot()
roc
```




```
augment(log_fit_titanic, new_data = titanic_test) %>%
  roc_auc(survived, .pred_Yes)
```

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.880
```

```
train_acc <- results %>% arrange(-accuracies)
test_acc <- augment(log_fit_titanic, new_data = titanic_test) %>%
  accuracy(truth = survived, estimate = .pred_class)
bind_cols(test_acc$.estimate, train_acc)
```

```
## New names:
## • `` -> `...1`
```

```
## # A tibble: 4 × 3
##    ...1 accuracies models
##   <dbl>      <dbl> <chr>
## 1 0.832      0.814 Logistic Regression
## 2 0.832      0.793 LDA
## 3 0.832      0.787 Naive Bayes
## 4 0.832      0.778 QDA
```

```
# We can see that the testing accuracy is a little higher than the training accuracy but
the difference is small. I guess maybe it's because the model has too few predictors so
 the model is underfitting.
```