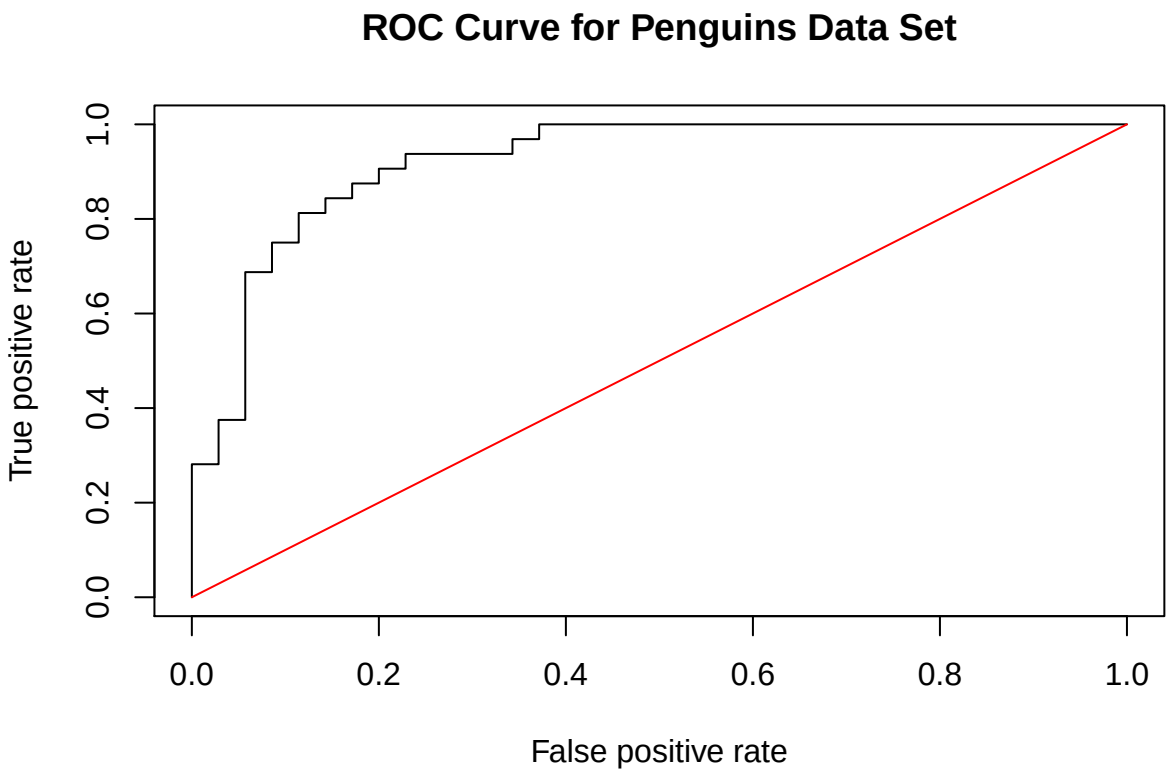
HW 12 Solutions

```
library(palmerpenguins)
library(tidyverse)
library(ROCR)

Data<-penguins

##remove penguins with gender missing
Data<-Data[complete.cases(Data[ , 7]),-c(2,8)]

##80-20 split
set.seed(1)
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
train<-Data[sample, ]
test<-Data[-sample, ]
```

1

(a)

```
##fit model
result<-glm(sex~species+bill_length_mm+bill_depth_mm+body_mass_g,
            family="binomial", data=train)

##generate ROC curve
preds<-predict(result,newdata=test, type="response")
rates<-prediction(preds, test$sex)
roc_result<-performance(rates,measure="tpr", x.measure="fpr")
plot(roc_result, main="ROC Curve for Penguins Data Set")
lines(x = c(0,1), y = c(0,1), col="red")
```



Since our ROC curve is closer to the top left than the red diagonal line, our logistic regression performs better than random guessing on the test data.

(b)

```
auc<-performance(rates, measure = "auc")
auc@y.values
```

```
## [[1]]
## [1] 0.9214286
```

Since the AUC is above 0.5, our logistic regression performs better than random guessing on the test data.

(c)

```
table(test$sex, preds>0.5)
```

```
##
##          FALSE TRUE
##   female    28    7
##   male       4   28
```

- The FPR is $\frac{7}{35} = 0.2$.
- The FNR is $\frac{4}{32} = 0.125$.
- The error rate is $\frac{7+4}{35+32} = 0.164$.

(d)

For this particular analysis, I do not believe there is a reason to favor reducing FPR over FNR (or vice versa); there isn't a worse consequence of wrongly identifying a female penguin as male versus wrongly identifying a male penguin as female. Thus we would prefer to reduce the overall error rate, which is achieved with a threshold of 0.5. So we do not need to adjust the threshold.