#Studying the survival rates from the titanic

#Setting the working directory

setwd("C:/Users/Ashley Rainbow/Desktop/Data Fellowship/data-fellowship/R Coding")

#Load Readr

library(readr)

#Loading the train data set

train <- read\_csv("train.csv")

View(train)

#Loading the test data set

test <- read\_csv("test.csv")

View(test)

#View the class of train data set

class(train)

#View the structure of the train data set

str(train)

#Set train$Pclass and train$Survived as a factor

train$Pclass <- as.factor(train$Pclass)

train$Survived <- as.factor(train$Survived)

#Set train$Name as a character

train$Name <- as.character(train$Name)

#Check for duplicate data

print("There should be 891 unique values.")

length(unique(train$PassengerId))

#View the structure of the test data set

str(test)

#Set test$Pclass as factor

test$Pclass <- as.factor(test$Pclass)

#Set test$Name as factor

test$Name <- as.character(test$Name)

#Check for duplicate data

print("There should be 418 unique values.")

length(unique(test$PassengerId))

#Explore the data

table(train$Survived)

#Store the values

train\_survived\_table <- table(train$Survived)

#Storing the count of survived and died

train\_count\_died <- c(train\_survived\_table[1])

train\_count\_survived <- c(train\_survived\_table[2])

#Creating proportions table

proportional\_survived <- prop.table(table(train$Survived))

#Loading ggplot2

library(ggplot2)

#Plotting sex and survival

ggplot(train, aes(x = Sex, fill = factor(Survived))) +

geom\_bar(width = 0.5) +

xlab("Sex") +

ylab("Total count") +

labs(fill = "Survived")

#Adding a model prediction column

train$ModelPrediction <- as.factor("TBD")

head(train)

#\*\*\*[this code is inneficient]\*\*\*

#model\_prediction <- rep(c("TBD"), each = 891)

#train <- cbind(train, model\_prediction)

#colnames(train)[which(names(train) == "model\_prediction")] <- "ModelPrediction"

#head(train)

#\*\*\*resume normal coding\*\*\*

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

class(train$ModelPrediction)

#Make all values in train$ModelPrediction "0"

train$ModelPrediction <- 0

head(train)

factor(train$ModelPrediction, levels = 0, 1)

head(train)

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Finding the percentage accuracy of prediction when all passengers do not survive

no\_survival <- sum(train$ModelPrediction == train$Survived)/nrow(train)

#Setting male prediction as 0

train$ModelPrediction[train$Sex == "male"] <- 0

#Setting female prediction as 1

train$ModelPrediction[train$Sex == "female"] <- 1

head(train)

#Finding the percentage accuracy of the prediction with females surviving

all\_females\_survive <- sum(train$ModelPrediction == train$Survived)/nrow(train)

#Creating a table for train classes

train\_passenger\_class <- table(train$Pclass)

#Assigning class counts to variables

train\_count\_first\_class <- c(train\_passenger\_class[1])

train\_count\_second\_class <- c(train\_passenger\_class[2])

train\_count\_third\_class <- c(train\_passenger\_class[3])

#Creating proportions table for passenger class

proportional\_class <- prop.table(table(train$Pclass, train$Survived))

proportional\_class

#Plotting gender survival rates including class.

ggplot(train, aes(x = Sex, fill = Survived)) +

geom\_bar(width = 0.5) +

facet\_wrap(~Pclass) +

ggtitle("Pclass") +

xlab("Sex") +

ylab("Total count") +

labs(fill = "Survived")

#Setting train$Modelprediction as 0

train$ModelPrediction <- 0

head(train)

#Setting train$Modelprediction as 1 for females in first and second class

train$ModelPrediction[train$Sex == "female"] <- 1

train$ModelPrediction[train$Pclass == "3"] <- 0

head(train)

##Finding the percentage accuracy of prediction when 1st and 2nd class females survived

first\_second\_class\_females\_accuracy <- sum(train$ModelPrediction == train$Survived)/nrow(train)

paste("The first model returned an accuracy of", all\_females\_survive, "the second model returned an accuracy of", first\_second\_class\_females\_accuracy)

#Plotting Pclass, Sex, Age and Survival

ggplot(train, aes(x = Age, fill = Survived)) +

facet\_wrap(~Sex + Pclass) +

geom\_bar(width = 10) +

xlab("Age") +

ylab("Total count")

#Resetting the train$ModelPrediction to 0

train$ModelPrediction <- 0

#Setting train$Modelprediction as 1 for females in first class, second class and for U18's

train$ModelPrediction[train$Sex == "female"] <- 1

train$ModelPrediction[train$Pclass == "3"] <- 0

train$ModelPrediction[(train$Pclass == "1" | train$Pclass == "2") & train$Sex == "male" & train$Age < 18] <- 1

#Finding the percentage accuracy of prediction when 1st and 2nd class females survived

upper\_class\_female\_and\_u18\_accuracy <- sum(train$ModelPrediction == train$Survived)/nrow(train)

paste("The first model returned an accuracy of", all\_females\_survive, "the second model returned an accuracy of", upper\_class\_female\_and\_u18\_accuracy)

#Creating submissions df

submission <- data.frame(test$PassengerId)

#Renaming the submission$test.PassengerId to PassengerId

names(submission)[names(submission) == "test.PassengerId"] <- "PassengerId"

View(submission)

#Setting survival rates to 0

submission$Survived <- as.integer(0)

factor(submission$Survived, levels = 1)

#Setting prediction values

submission$Survived[test$Sex == "female"] <- 1

submission$Survived[test$Pclass == "3"] <- 0

submission$Survived[(test$Pclass == "1" | test$Pclass == "2") & test$Sex == "male" & test$Age < 18] <- 1

View(submission)

#Export submission

write.csv(submission, file = "AshleyRainbow\_Virtualstock\_TitanicSubmission.csv")





