STAT GR 4206/5206 Homework 1

Christine Chong cc4190

May 23, 2017

Part One

i. Import the titanic data set into RStudio using read.table(). Use the argument as.is = TRUE. The data set should be stored in a data frame called titanic.

We can use **read.table** to import the data from **Titanic.txt**

```
titanictable <- read.table("Titanic.txt", header=TRUE, as.is=TRUE)
```

ii. How many rows and columns does titanic have?

We can use **nrow** and **ncol** function to find the rows and columns

```
nrow(titanictable)
```

```
## [1] 891
```

```
ncol(titanictable)
```

```
## [1] 12
```

iii. Create a new variable in the data frame called Survived.Word. It should read either "survived" or "died" indicating whether the passenger survived or died. This variable should be of type "character"

We can use **ifelse** to separate the passengers who survived from those who did not using the labels **Survived** and **Dead**. Then we can assign this **ifelse** statement into a vector called **Survived.Word** so that we can use the vector to reference the data from the **ifelse** statement. Then we can create a new column the table **titanictable** using **$**.

```
Survived.Word <- ifelse(titanictable$Survived == 1, "Survived" , "Died")
titanictable$Survived.Word <- Survived.Word
```

Part 2: Exploring the Data in R

i. Use the apply() function to calculate the mean of the variables Survived, Age, and Fare. This will require using the apply() function on a sub-matrix of dimension 891*3. Explain what the mean of Survived tells us. One of the mean values is NA. Which variable has a mean value of NA and why is this the case?

We can assign variables to the columns of survived, age and fare. Then we can assign a vector called **submatrix** that contains all the values of these variables. Then we can use the **matrix** function using the **submatrix** as the data. Then we can use the **apply** function in order to find the means of all of the columns.

```
s <- titanictable$Survived
a <- titanictable$Age
f <- titanictable$Fare
submatrixdat <- c(s, a, f)
submatrix <- matrix(data = submatrixdat, nrow = 891, ncol = 3, byrow = FALSE)
apply(submatrix, 2, mean)
```

```
## [1]  0.3838384         NA 32.2042080
```

The mean of those who survived (0.3838) indicates to us that around 38.38% of people were categorized as survived from the accident.

The reason why the mean of the age of the passengers is NA because there was age data for passengers that were recorded as NA. As discussed in class, if one of the numbers of a data is NA, then the mean will be NA because there is no way to take the mean between numbers and character strings for R. (Unless if the NA data is excluded).

ii. Compute the proportion of female passengers who survived the titanic disaster. Round your answer to 2 decimals using the round() function.

First we should create a subset of the data that contains data on only the passengers who survived in order to get a count of the amount of females that have survived. Then we create a subset of data of all the female passengers in order to get a total on the number of female passengers. Then we can use this number and compare it to the number of all female passengers.

```
subMatrixAllFemale <- subset(titanictable, titanictable$Sex == "female")
subMatrixSurvived <- subset(titanictable, titanictable$Survived == 1)
gender <- subMatrixSurvived$Sex
submatrixgenderdat <- c(gender)
submatrixgender <- matrix(data = submatrixgenderdat,  byrow = FALSE)
female <- table(submatrixgender)["female"]
allfemale <- nrow(subMatrixAllFemale)
fToAllF <- female / allfemale
round(fToAllF, digits = 2)
```

```
## female
##   0.74
```

iii. Of the survivors, compute the proportion of female passengers. Round your answer to 2 decimals. This answer may take a few lines of code. One strategy would be to create a survivors matrix that only includes individuals who survived the disaster. Then using the survived matrix, calculate the proportion of females.

Using the earlier matrix, we can simply use the numbers of all female passengers and the number of all passengers to calculate the proportion of female passengers to male passengers.

```
male <- table(submatrixgender)["male"]
fToAllS <- female/(male + female)
round(fToAllS, digits = 2)
```

```
## female
##   0.68
```

iv. Use the following code to create an empty numeric vector of length three called Pclass.Survival. We will fill in the elements of Pclass.Survival with the survival rates of the three classes.

```
classes <- sort(unique(titanictable$Pclass))
Pclass.Survival <- vector("numeric", length = 3)
names(Pclass.Survival) <- classes
```

Next use a for loop to fill in Pclass.Survival vector with the survival rates for each class. The statements inside the loop should update the vector Pclass.Survival with the survival rate (the proportion of people who survived for each class. Your loop should look like the following, with of course, your own code added inside the loop). The elements in the Pclass.Survival vector should be rounded to two decimal places.

We can first create a matrix that is only concerned about the class for those who survived. Then we can use the **for loop** to insert the number of survivors for each class into the **Pclass.Survival** vector and divide that number by all the survivors in order to get the rate of survival for each class.

```
shipclass <- subMatrixSurvived$Pclass
submatrixclassdat <- c(shipclass)
submatrixclass <- matrix (data = submatrixclassdat, byrow = FALSE)
allsurvivors = length(submatrixclass)
for(i in 1:3){
  Pclass.Survival[i] <- round(table(submatrixclass)[i] / allsurvivors, digits = 2)
}
Pclass.Survival
```

```
##    1    2    3
## 0.40 0.25 0.35
```

v. Now create a Pclass.Survival2 vector that should equal the Pclass.Survival vector from the previous question, but use the tapply() function. Again, round the values to 2 decimals

We can use the **tapply** function using the Survived column (to count the amount of survivors), and Pclass column (to categorize them into groups by class) of the subMatrixSurvived subset and a custom function that calculates proportion by dividing the length of each class and the number of all survivors. To round the value to 2 decimals we can use the **round** function.

```
Pclass.Survival2 <- vector("numeric", length = 3)
Pclass.Survival2 <- tapply(subMatrixSurvived$Survived,
                           subMatrixSurvived$Pclass,
                           function(x){ round(length(x) / allsurvivors, digits =2 )})
Pclass.Survival2
```

```
##    1    2    3
## 0.40 0.25 0.35
```

vi. Does there appear to be a relationship between survival rate and class?

There does not appear to be a relationship between survival rate and class, however the highest proportion of survivors was from the first class.