HW1 - Titanic

Salim M'jahad msm2243

February 6, 2018

Part 1: Importing Data into R

i. Importing the Titanic dataset into RStudio

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

ii. Number of rows and columns of data

```
dim(titanic)
```

```
## [1] 891  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.

```
Survived.Word <- titanic$Survived
Survived.Word[Survived.Word == 1] <- "survived"
Survived.Word[Survived.Word == 0] <- "died"
typeof(Survived.Word)
```

```
## [1] "character"
```

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?

```
three <- cbind(titanic["Survived"],titanic["Age"],titanic["Fare"])
apply(three, 2, mean)
```

```
##   Survived        Age       Fare
##  0.3838384         NA 32.2042080
```

Age has a mean of NA because the Age column has some NA values. R does not skip those values by default.

ii. Compute the proportion of female passengers who survived the titanic disaster. Round your answer to 2 decimals using the *round()* function. Hint *?round*. Note: This is not a conditional probability.

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. Note: This is a conditional probability

```
survivors <- subset(titanic, Survived==1)$Sex
print(round(length(subset(survivors, survivors=="female"))/length(titanic$Survived), digits=2))
```

```
## [1] 0.26
```

```r
print(round(length(subset(survivors, survivors=="female"))/length(survivors), digits=2))
```

```
## [1] 0.68
```

26% of passengers are female survivors. 68% of survivors are female.

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.

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

for (i in 1:3) {
    class_sub <- subset(titanic, Pclass==i)
    surv <- length(subset(class_sub$Survived, class_sub$Survived==1))/length(class_sub$Survived)
    Pclass.Survival[i] <- round(surv, digits = 2)

}
Pclass.Survival
```

```
##    1    2    3
## 0.63 0.47 0.24
```

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.

```r
Pclass.Survival2 <- round(tapply(titanic$Survived, titanic$Pclass, mean), digits=2)
Pclass.Survival2
```

```
##    1    2    3
## 0.63 0.47 0.24
```

vi. The more premium a passenger's class (closer to 1 aka lower although it is a "higher class"), the more likely they are to survive the Titanic crash.