Homework 1 Solutions

Part 1

i. Looking at *Titanic.txt* I see that the data is tab delimited and therefore I use *read.table()*. Since the first row of data in the text file holds the column names, I use *header = TRUE*.

```
setwd("~/Desktop/Data")
titanic <- read.table("Titanic.txt", header = TRUE, as.is = TRUE)
```

ii. The function *dim()* provides the dimension of its input object.

```
dim(titanic)
```

```
## [1] 891  12
```

```
str(titanic)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

iii. There are multiple ways to do this. In the following, I add a new column called *Survived.Word* with each entry equal to *"survived"*. Then I reassign the values in the variable *Survived.Word* to *"died"* in the rows where *Survived* equals 0.

```
titanic$Survived.Word <- "survived"
titanic$Survived.Word[titanic$Survived == 0] <- "died"
```

Alternatively, we could use *ifelse()* to complete this. The first argument is a logical vector that's *TRUE* when *Survived* equals 1 and *FALSE* when *Survived* equals 0.

```
titanic$Survived.Word <- ifelse(titanic$Survived == 1, "surived", "died")
```

Part 2

i. To solve this problem we create a sub-matrix of the variables of entry and then pass this sub-matrix into the *apply()* command.

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

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

The mean of *"Survived"* is the *proportion* of passengers that survived the disaster. The mean of the *Age* variable is *NA*, because some of the passenger's ages are unknown (i.e. *Age* also has some missing values).

ii. As in the last question, we can calculate the proportion of survivors by taking the mean of the *Survived* variable, but here we filter to only include female passengers.

```
round(mean(titanic$Survived[titanic$Sex == "female"]), 2)
```

```
## [1] 0.74
```

iii. To answer this question we create a sub-matrix *survivors* which only includes the rows of *titanic* corresponding to those surviving the disater. Then we calculate the proportion as the number of female passengers in the *survivors* matrix divided by the total number of people in the *survivors* matrix.

```
survivors     <- titanic[titanic$Survived == 1, ]
proportion    <- sum(survivors$Sex == "female")/length(survivors$Sex)
round(proportion, 2)
```

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

Alternatively, we can use the *table()* command.

```
survivors     <- titanic[titanic$Survived == 1, ]
proportion    <- table(survivors$Sex)/length(survivors$Sex)[1]
round(proportion, 2)
```

```
##
## female   male
##   0.68   0.32
```

iv.

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

for (i in 1:3) {
  thisclass          <- titanic[titanic$Pclass == i, ]
  Pclass.Survival[i] <- round(mean(thisclass$Survived), 2)
}
```

v.

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

```
##    1    2    3
## TRUE TRUE TRUE
```

vi.

```
Pclass.Survival
```

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

There does appear to be a relationship between survival and class. We can see from the previous question that the survival rate decreases with ticket class, meaning fewer members of the lower class survived than members of the upper class.