# Assignment1

Giang Vu

1/29/2021

## Assignment 1

### Part 1

**i.** Data was imported

```
#set working dir.
setwd("/Users/giangvu/Desktop/STAT 2301 - Statistical Computing and Data Science/HW/HW1")
#import dataset
titanic <- read.table("Titanic.txt",header = T,as.is = T)
```

**ii.** There are 12 columns and 891 rows in the dataframe.

```
#number of columns
ncol(titanic)
```

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

```
#number of rows
nrow(titanic)
```

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

**iii.** A new variable "Survived.Word" was created below.

```
#create new column Survived.Word
titanic$Survived.Word <- ifelse(titanic$Survived == 0, "died", "survived")
#check data to see if it worked
head(titanic[,c("Survived","Survived.Word")])
```

```
##   Survived Survived.Word
## 1        0          died
## 2        1      survived
## 3        1      survived
## 4        1      survived
## 5        0          died
## 6        0          died
```

**Part 2**

**i.** The mean of "Survived" is around 0.38, meaning on average 38% of the people on the Titanic survived. The mean of "Age" is "NA" because there are many missing values ("NA") in this column, and we didn't remove those observations with missing values so R gave us a result of "NA" when we tried to do calculations on such column.

```
#create a 891x3 matrix of Survived, Age, and Fare
mtrx1 <- matrix(nrow = 891,ncol = 3)
colnames(mtrx1) <- c("Survived","Age","Fare")
mtrx1[,1] <- titanic$Survived
mtrx1[,2] <- titanic$Age
mtrx1[,3] <- titanic$Fare

#calculate column mean of this new matrix
apply(mtrx1, 2, mean)
```

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

**ii.** Proportion of female passengers who survived was 0.26.

```
#proportion of female passengers who survived
round(nrow(titanic %>% filter(Survived==1 & Sex=="female"))/nrow(titanic),2)
```

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

**iii.** Out of the survivors, the proportion of female passengers was 0.68.

```
#extract only survivors
survivors <- titanic %>% filter(Survived==1)

#proportion of female in survivors
round(nrow(survivors %>% filter(Sex=="female"))/nrow(survivors),2)
```

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

**iv.** The vector Pclass.Survival is created as follows.

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

#fill in the empty vector with survival rate for each class
for (i in 1:3) {
  srate <- round(nrow(titanic %>%
                        filter(Pclass==i & Survived==1))/nrow(titanic %>% filter(Pclass==i)),2)
  Pclass.Survival[i] <- srate
}
```

```
#result
Pclass.Survival
```

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

**v.** The vector Pclass.Survival2 is created as follows.

```
#create empty vector
Pclass.Survival2 <- vector("numeric", length = 3)
names(Pclass.Survival2) <- classes

#create vectors for tapply from data
classes.fct <- factor(titanic$Pclass) #a factor of passenger class for all passengers
surv.vct <- as.vector(titanic$Survived) #a vector of survival status for all passengers

#fill in the empty vector with survival rate for each class
Pclass.Survival2 <- round(tapply(surv.vct, classes.fct, mean),2)
#because the mean of survival status, which takes on values 0 and 1, is the survival rate/proportion.

#result
Pclass.Survival2
```

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

**vi.** There seems to be a correlation here between survival rate and class. As class decreased from 1st to 2nd to 3rd, the survival rate lowered. Higher (more luxury) passenger class had a higher survival rate.