Hwk1_bs2996

Bangda Sun

September 22, 2016

Part 1: Importing Data into R

Question (i)

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

```
data <- read.table("C:\\Users\\Bangda\\Desktop\\GR5206 Materials\\Titanic.txt",
                   header = T, sep = "", as.is = T)
titanic <- data.frame(data)
is.data.frame(titanic) # Check if titanic is a data.frame
```

```
## [1] TRUE
```

Question (ii)

How many rows and columns does **titanic** have? (If there are not 891 rows and 12 columns something is wrong. Check part (i) to see what could have gone wrong.)

```
rows <- dim(titanic)[1]; rows
```

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

```
cols <- dim(titanic)[2]; cols
```

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

Question (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`.

```
titanic$Survived.Word[titanic$Survived == 1] <- "survived"
titanic$Survived.Word[titanic$Survived == 0] <- "died"
is.character(titanic$Survived.Word) # Check if the new variable is character
```

```
## [1] TRUE
```

```
table(titanic$Survived, titanic$Survived.Word) # Check the results of assignment
```

```
##
##     died survived
##   0  549        0
##   1    0      342
```

Part 2: Exploring the Data in R

Question (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?

```
age <- titanic$Age
fare <- titanic$Fare
survived <- titanic$Survived
submat <- cbind(survived, age, fare)
apply(submat, 2, mean)
```

```
##   survived        age       fare
##  0.3838384         NA 32.2042080
```

Variable **Age** has a mean of value of `NA` because there are some missing values in **Age**.

Question (ii)

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

```
group_sex <- factor(titanic$Sex)
survived_list <- split(titanic$Survived, group_sex)
prop <- mean(survived_list$female)
prop <- round(prop, digits = 2); prop
```

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

Question (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.

```
num_male_survived <- sum(survived_list$male)
num_female_survived <- sum(survived_list$female)
prop2 <- num_female_survived / (num_male_survived + num_female_survived)
prop2 <- round(prop2, digits = 2); prop2
```

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

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

Next use a `for` loop to fill in the **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.

```
for (i in 1:3) {
code that fills in the Pclass.Survival vector
}
```

The elements in the **Pclass.Survival** vector should be rounded to two decimal places.

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

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

Question (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.

```
group_Pclass <- factor(titanic$Pclass)
survived_rate <- tapply(titanic$Survived, group_Pclass, mean)
survived_rate <- round(survived_rate, digits = 2); survived_rate
```

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

Quesion (vi)

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

Answer: Yes. Survival rate decrease as the number of class increase.