STAT GR4206/5206 Homework 1 [100 pts]
Due 8:00pm Tuesday, May 30th on Canvas

Your homework should be submitted on Canvas as an R Markdown file. Please submit both the .Rmd and knitted .pdf files. We will not accept any other formats. Please clearly label the questions in your responses and support your answers by textual explanations and the code you use to produce the result. Note that you cannot answer the questions by observing the data in the "Environment" section of RStudio or in Excel – you must use coded commands.

Problem 1

In this assignment we'll be studying a data set which provides information on the survival rates of passengers on the fatal voyage of the ocean liner *Titanic*. The dataset provides information on each passenger including, for example, economic status, sex, age, cabin, name, and survival status. This is a training dataset taken from the Kaggle competition website; for more information on Kaggle competitions, please refer to `https://www.kaggle.com`. Students should download the data set on Canvas. Below is a more detailed description of the variables.

Table 1: VARIABLE DESCRIPTIONS

`PassengerId`	Identification Number
`Survived`	Survival (0 = No; 1 = Yes)
`Pclass`	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
`Name`	Name
`Sex`	Sex
`Age`	Age
`SibSp`	Number of Siblings/Spouses Aboard
`Parch`	Number of Parents/Children Aboard
`Ticket`	Ticket Number
`Fare`	Passenger Fare
`Cabin`	Cabin
`Embarked`	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Perform the following tasks:

Part 1: Importing Data into `R`

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

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.)

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'.

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?

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

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.

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.

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.

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