**Lab Assignment #3 (Due: 11:00am, Tue, Sep 4)**

If you cannot finish it by 11am, please talk to me. I’ll give you an extension. No Penalty!!

**Write Your Name Here**

**<< Attention >>** You can use this file for our lab assignment. Please edit as you wish. (You should delete unnecessary parts and add your work.)

**1. The Titanic (Data:** **Titanic)**

On April 15, 1912, on her maiden voyage, the Titanic collides with an iceberg and sank. The ship was luxurious but did not have enough lifeboats for the 2,224 passengers and crew. As a result of the collision, 1,502 people died. The ship had three classes of passengers. The level of luxury and the price of the ticket varied with the class, with first class being the most luxurious.

* First, import the data file “**Titanic.csv**” into RStudio.
* Then, attach the data file

> attach(Titanic)

**(a)** Find the frequencies.

* Paste the R command and the result inside a text box.
* Complete the table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Ticket Class** | | |
| **First Class** | **Second Class** | **Third Class** |
| **Status** | **Died** |  |  |  |
| **Survived** |  |  |  |

**(b)** Let’s compare the death/survival rates of passengers in each ticket class (i.e. relationship between the passenger’s survival status and the passenger’s ticket class).

* Draw a side-by-side bar plot using passenger class and survival status. The category names of ticket class (First, Second, Third) should appear in the horizontal axis. **Note:** See “R Handout2” for help.
* Add colors in the plot.
* Add a legend at the top-left corner of the plot.
* Paste the R commands inside a text box.
* Paste the side-by-side bar plot here. No need to use a text box, Resize the plot if it is too large or too small.

**(c)** Let’s take a look at the distribution of passengers’ ages using a histogram.

* Draw a histogram of passengers’ ages.
* Paste the R command inside a text box.
* Paste the histogram here. No need to use a text box, Resize the plot if it is too large or too small.

**(d)** Calculate the mean age and the median age of thepassengers. Also, find the age of the youngest and the oldest passengers. **Note:** See “R Handout1” for help.

* Paste the R command and the results here inside a text box. **Hint:** You can type summary(Age) to calculate the sample statistics (5-number summary and mean).

**(e)** Answer it based on the results in (c) and (d). Write sentences.

* Describe the distributional shape: Symmetric, Right-Skewed, or Left-Skewed.
* Are there more young passengers or older passengers?
* What are the ages of the youngest and the oldest passengers?
* What is the typical age of passengers?

**(f)** Let’s compare the typical age of passengers in each ticket class using histograms.

* Type the following commands.

> Age.First = Age[ Class=="First" ]

> Age.Second = Age[ Class=="Second" ]

> Age.Third = Age[ Class=="Third" ]

* Age.First will contain ages of the first-class passengers only
* Age.Second will contain ages of the second-class passengers only
* Age.Third will contain ages of the third-class passengers only
* Draw a histogram of ages in each ticket class. We’ll use the same x-axis. Type the following commands.

> hist(Age.First, xlim=c(0, 80))

> hist(Age.Second, xlim=c(0, 80))

> hist(Age.Third, xlim=c(0, 80))

* Paste the R commands inside a text box.
* Paste the histograms here. No need to use a text box. It would be nice to have three histograms side-by-side.

**(g)** Let’s compare the typical age of passengers in each ticket class.

* Calculate the five-number summary in each ticket class. Type the following

> tapply(Age, Class, summary)

.

* Copy the R commands and the result inside a text box.

**(h)** Answer it based on the results in (f) and (g). Write sentences.

* Describe the distributional shape (Symmetric, Right-Skewed, or Left-Skewed) of each histogram.
* What is the typical age of passengers in each ticket class? Do you think that passengers of a particular class are older (or younger) than passengers in the other classes? Explain why or why not. **Note**: We’ll learn whether there is a significance difference in ages in the later chapter.

**(i)** Let’s compare the typical ages of passengers who survived and passengers who died.

* Copy the R commands and the result inside a text box.
* What is the typical age of passengers who died? What is the typical age of passengers who survived? Do you think that survived passengers are younger or older than the passengers who died? Explain why or why not.

**2. Impact of Using Celebrities (Data:** **Commercial**)

Consider a study in the Journal of Marketing on the impact of using celebrities in television advertisements. They randomly selected TV viewers and ask them to watch a commercial (e.g. Super Bowl Commercial - Make It with Keanu Reeves <https://www.youtube.com/watch?v=W3FH1scDhfA>)

Then, each person is asked to identify the product. The researchers investigated the relationship between the gender of a viewer and the viewer’s brand awareness.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Identify | Whether a person can identify the brand or not (Yes or No) |
| Gender | Gender of a Viewer (Male or Female) |

* First, import the data file “**Commercial.csv**” into RStudio.
* Then, attach the data file

> attach(Commercial)

The following object is masked from Titanic:

Gender

**Attention**: If you have attached the “Titanic” data in the same window, you’ll see the error message as shown above. It happens because two files have the same variable “Gender”. It is the best practice to detach the data file detach(Titanic) once you are done. Then, you can attach a different data file.

**Note**: In the future, you can avoid this problem by

* Detach the file when you finished.
* Or you can start a new RStudio to work with a different data file.

**(a)** We have two categorical variables. Find the frequencies.

* Paste the R command and the result inside a text box.
* Complete the table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Brand Awareness** | |
| **Can Identify** | **Cannot Identify** |
| **Gender** | **Female** |  |  |
| **Male** |  |  |

**(b)** Let’s investigate the relationship between the gender of a viewer and the viewer’s brand awareness.

* Draw a side-by-side bar plot. The category names of gender (Male and Female) should appear in the horizontal axis.
* Add colors in the plot.
* Add a legend in the plot.
* Paste the R commands inside a text box.
* Paste the side-by-side bar plot here.

**(c)** Discuss the relationship between the gender of a viewer and the viewer’s brand awareness using the results in (a) and (b).